

# **Ethical Regulators and Super-Ethical Systems**

**Mick Ashby**  
ethics@ashby.de

## **Abstract**

This paper takes the Good Regulator Theorem, and unifies it with the Law of Requisite Variety and seven other requisites that are necessary and sufficient for a cybernetic regulator to be effective and ethical. The resulting Ethical Regulator Theorem provides a basis for systematically evaluating the adequacy of existing or proposed designs for systems that make decisions that can have ethical consequences; regardless of whether the regulators are human, machines, or cyberanthropic hybrids. The theorem has potentially far-reaching implications for society. A new framework is proposed for classifying cybernetic systems, which highlights the existence of a possibility-space bifurcation in our future time-line, and the implementation of “super-ethical” systems is identified as an urgent moral imperative for the human race to avoid a technological dystopia. Concrete actions are proposed to steer our future towards a cyberanthropic utopia.

Keywords: Ethics, Transparency, Robotics, Singularity, Cyberanthropic Utopia

## **The Ethical Regulator Theorem**

The Good Regulator Theorem (Conant 1970) is ambiguous because a regulator that is good at regulating is not necessarily good in an ethical sense. To avoid this ambiguity, this paper uses the term “effective” for the first meaning, “ethical” for the second, and only uses “good” when both meanings are intended.

The Good Regulator Theorem proved that every effective regulator of a system must be a model of that system, and the Law of Requisite Variety (Ashby 1956) dictates the range of responses that an effective regulator must be capable of. However, having an internal model and a sufficient range of responses is insufficient to ensure effective regulation, let alone ethical regulation. And whereas being effective does not require being optimal, being ethical is absolute with respect to a particular ethical schema.

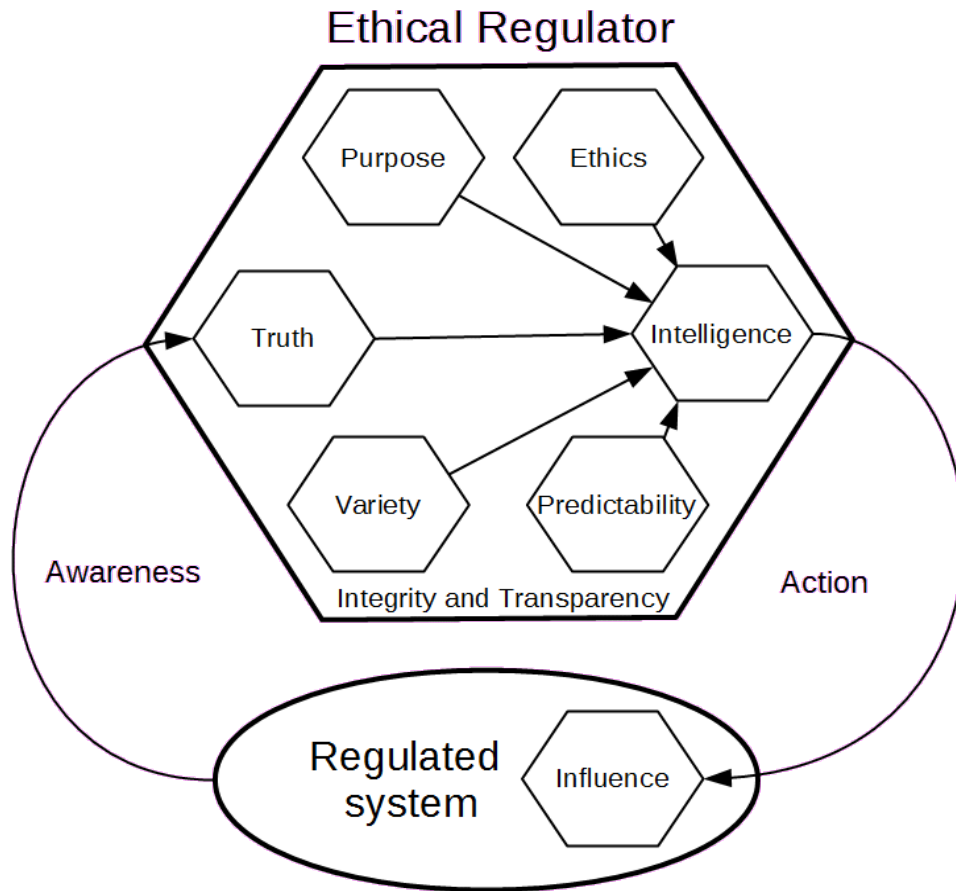
The Ethical Regulator Theorem claims that the following nine requisites are necessary and sufficient for a cybernetic regulator to be effective and ethical:

1. Truth about the past and present.
2. Variety of possible actions.
3. Predictability of the future effects of actions.
4. Purpose expressed as unambiguously prioritized goals.

## Ethical Regulators and Super-Ethical Systems

5. Ethics expressed as unambiguously prioritized rules.
6. Intelligence to choose the best actions.
7. Influence on the system being regulated.
8. Integrity of all subsystems.
9. Transparency of ethical behaviour.

Figure 1 and the following sections explain the requisites in more detail.



**Figure 1: The Ethical Regulator System**

### Requisite Truth

**Truth** is not just about information that the regulator treats as facts or receives as inputs, but also the reliability of any interpretations of such information. This is the regulator's awareness of the current situation and knowledge. If the regulator's information sources or interpretations are unreliable, and cannot be error-corrected, then the integrity of the system is in danger. In extremis, if the perceptions of the regulator can be manipulated, it can be tricked into making decisions that are ineffective or unethical.

## Ethical Regulators and Super-Ethical Systems

101-philosophers might claim that objective truth is unattainable, therefore requisite truth is unattainable, therefore no ethical regulator can exist. However, this argument is a fallacy. An ethical regulator doesn't require perfectly accurate information, rather it must be able to cope ethically with uncertainties and minimize the impact of unreliable information, misinterpretations, and deliberate misinformation as best as it can. This is much like the requirement that a good judge (effective and ethical) must be able to reach reliable verdicts "beyond reasonable doubt" from unreliable evidence.

### Requisite Variety

**Variety** in the range of possible actions must be as rich as the range of potential disturbances or situations. This is nothing other than the Law of Requisite Variety.

### Requisite Predictability

**Predictability** requires a sufficiently accurate model (of the system being regulated, including the regulator) that can be used to rank the actions that will give the best outcome. This is nothing other than the Good Regulator Theorem.

### Requisite Purpose

**Purpose** must be expressed as unambiguously prioritized goals, because complex systems are generally required to satisfy multiple goals. Without goals, the system cannot be effective. These goals cannot violate any ethical imperatives.

### Requisite Ethics

**Ethics** must be expressed as unambiguously prioritized values that codify rules, regulations, and laws in a human-readable form. Ethical goals have a higher priority than goals for purpose. By always obeying the relevant highest priority ethical imperatives, the regulator is guaranteed to act ethically within the scope of the ethical schema.

Because ethical schemas vary between different cultures, they must be handled as plug-ins. And because an ethical schema can encode any ethics, good or bad, each ethical schema must be anchored in the laws of a particular legislative jurisdiction.

For example, a self-driving vehicle must avoid harming people and animals. In the U.S.A. and Europe, protecting a person is a higher ethical priority than protecting a cow. But, in India, where cows are sacred, the opposite priorities might be regarded as ethically preferable. Therefore, crossing a state or national border might require activating a different ethical schema.

A taxonomy of ethics modules can provide ethical coverage for all conceivable ethical situations. For example, medical, first-aid, child-care, traffic-rules, gun-law, tax-law, contract-law, maritime-law, drone-flying, police-regulations, operating-nuclear-power-stations, and warfare-rules-of-engagement.

Ethics modules can be treated like ethical device drivers, so that to be fully operational, a hypothetical gun-carrying, tax-advising robot that can drive on roads requires valid ethics

## Ethical Regulators and Super-Ethical Systems

modules for gun-law, tax-code, and traffic-rules. Without all necessary modules for the appropriate legal jurisdiction, the robot's gun, tax advising, or driving capabilities are automatically disabled.

By legislating that all autonomous artificial intelligence (AI) systems must include and obey appropriate ethics modules that are issued by an organization that is run by humans, we can establish a control mechanism that should ensure that intelligent machines are always subject to human ethics; including Isaac Asimov's Three Laws of Robotics (Asimov 1942); without unduly restricting the freedom of AI researchers. In fact, it will free AI researchers and knowledge engineers to focus on the more challenging requisites of truth, predictability, and intelligence.

### Requisite Intelligence

**Intelligence** must be applied to the previous five requisite types of information to select the most rational and effective ethical action from the set of possible actions.

### Requisite Influence

**Influence** is the existence of pathways to transmit the effects of the selected actions to the regulated system. This is not a property of the regulator itself, but a function of the connectivity relationships that span from the regulator's outputs to elements of the regulated system and its environment.

A regulator that is isolated from influencing the regulated system is not a true regulator, it is just an observer or a simulation. As an observer or simulation, there are no direct ethical consequences; which can be important when observing or simulating dangerous situations.

The immediacy of the effect of actions can vary greatly, depending on the nature of the system being regulated. For example, the Supreme Court issuing a ruling; a self-driving vehicle applying the brakes; or someone sending a message to a complex network of amplifying and variable-delay transmission repeaters, known as Twitter followers.

In some systems, influence is more of a determining factor than variety. Indeed, the power of the Law of Requisite Variety has often been overstated, for example, claiming that the subsystem with the most variety will control a system. This is not always true.

If we consider where two systems, A and B, are competing to win control of system C, for example, two politicians seeking election, often the variety of statements, actions, and strategies of the candidates is less important than their ability to purchase advertising to influence the voters.

And if a robber uses a gun to increase his chances of success, the use of a gun does not amplify his variety, it is just one existing element in his range of variety, yet making that choice greatly increases his effectiveness at controlling his victims. Such an increase in effectiveness, like buying advertising, is best explained in terms of an increase in influence. The variety of the robber or an advertising message is effectively constant.

## Ethical Regulators and Super-Ethical Systems

### *Effectiveness Function*

We can define a pseudo-metric effectiveness function that expresses how the effectiveness that a regulator, R, has in controlling a system to achieve a given goal actually depends on the quality or strength of five requisites.

$$\text{Effectiveness}_R = \text{Truth}_R \times \text{Variety}_R \times \text{Predictability}_R \times \text{Intelligence}_R \times \text{Influence}_R$$

So if  $\text{Effectiveness}_A$  is greater than  $\text{Effectiveness}_B$ , then A is more likely than B to win control over system C. And if the quality or strength of  $\text{Truth}_A$ ,  $\text{Variety}_A$ ,  $\text{Predictability}_A$ ,  $\text{Intelligence}_A$ , or  $\text{Influence}_A$  gets close to zero, the effectiveness of A is massively reduced.

When we introduce ethics, the effectiveness formula must be modified because the effect of behaving ethically is that it reduces the variety of options that are available by removing all possibilities that are unethical. Thus if A is an ethical politician, and B is an unethical politician, we get:

$$\begin{aligned} \text{Effectiveness}_A &= \text{Truth}_A \times (\text{Variety}_A - \text{Ethics}_A) \times \text{Predictability}_A \times \text{Intelligence}_A \times \text{Influence}_A \\ \text{Effectiveness}_B &= \text{Truth}_B \times \text{Variety}_B \times \text{Predictability}_B \times \text{Intelligence}_B \times \text{Influence}_B \end{aligned}$$

Which captures the reality that, with all other things being equal, businessmen and politicians who lie and cheat have an advantage over ones that are ethical.

It is worth noting that in social systems, there is no upper-limit on the potential power of money to purchase media influence; which in turn, if the media is broadcasting advertising, lies, or propaganda, reduces the quality of  $\text{Truth}_x$  that is received by every consumer or voter, X, which can manipulate them into making decisions that are not in their best interest.

Although the effectiveness function is stated pseudo-mathematically, it is neither necessary, nor possible to calculate meaningful numerical values in order to compare the effectiveness of different systems or configurations. The essential value of the function is to understand the relationships and dependencies that it captures. It is sufficient if an intuitive understanding of the effectiveness function informs the system design strategy; recognizing that a maximally effective and ethical system requires that the quality of five requisite dimensions are maximized and that a successful attack on the integrity of any one of the nine requisites spells disaster for the effectiveness and/or ethical adequacy of the whole system.

### Requisite Integrity

**Integrity** of the regulator and all its subsystems must be assured through features, such as resistance to tampering, intrusion detection, and cryptographically authenticated ethics modules. Monitoring mechanisms must identify if any invalid ethics modules are being used or if an ethical imperative is violated, and if necessary, automatically notify the appropriate authorities, preserve evidence, and activate an ethical fail-safe mode.

Note that the regulator's influence is outside the scope of its first-order integrity mechanisms, which poses a potential vulnerability that can only be mitigated by using the

## Ethical Regulators and Super-Ethical Systems

situation awareness closed-loop feedback to check for evidence of the effect of each action. For example, you can be confident that your tweet was received and distributed by Twitter when you see that it has been retweeted.

### Requisite Transparency

**Transparency** is defined by introducing **The Law of Ethical Transparency**, which states “For a system to be truly ethical, it must be possible to prove retrospectively that it acted ethically with respect to the appropriate ethical schema.”.

Whereas it really doesn’t matter whether the programmers of a chess playing robot can find out why a piece was sacrificed, the logic of ethical decisions must not be hidden in the depths of opaque processes or lost to the passage of time. Generally, this requisite can only be satisfied by keeping audit trails that are adequate and secure.

When an ethically adequate system violates an ethical imperative, as they sometimes will, analysis of the audit trail will identify the reason. For example, because a boy leading a cow was mistakenly identified as a calf leading a man, or it will prove whether a CEO was lied to about illegal corporate activities.

**Integrity** and **Transparency** are codependent because we require both integrity of transparency and transparency of integrity.

### Evaluating Ethical Adequacy

The evaluation of ethical adequacy has strong similarities to network penetration testing, where the evaluator tries to identify theoretical possibilities to subvert the integrity of the system.

An evaluated system is judged on the adequacy of each requisite dimension. Only systems that meet all nine requisites can be said to be “ethically adequate”. Systems that do not fulfil all nine requisites are classified as “ethically inadequate” and the weaknesses listed with recommendations for improving them.

Perhaps, in the near future, accredited ethical consultants will specialize in auditing and certifying the ethical adequacy of existing and proposed systems and processes.

This theorem cannot be used to certify that an ethical schema is ethical because schemas can vary arbitrarily between different cultures. However, existing automatic proof algorithms should be able to detect certain types of errors before an ethical schema is packaged as a module.

### The Law of Inevitable Ethical Inadequacy

This new law states: “If you don’t specify that you require a secure ethical system, what you get is an insecure unethical system.”.

## Ethical Regulators and Super-Ethical Systems

The reason is because when ethical adequacy is not a requirement for a system design, the resulting design will tend to optimize for effectiveness and therefore maximally ignore the ethical, integrity, and transparency dimensions, which are optional for a system that only needs to be effective, thus guaranteeing that any implementation will be ethically inadequate and vulnerable to manipulation; by design.

### Legislative Implications

By creating a well-defined interface for coding ethics, it becomes easier to apportion legal liability for failures. For example, if a self-driving car crosses the border into India, fails to switch to the Indian government certified ethics module for traffic-rules, and decides to hit a cow to avoid hitting a person, then the car manufacturer can be held liable for the crime of killing a sacred animal. But if the correct ethics module was activated, but the “don’t hit cows” rule had an incorrectly low priority in the ethics schema, then the provider of the ethics module might be liable.

It is only a matter of time until the laws and regulations of every country are available in a standardized XML format such as LKIF (Legal Knowledge Interchange Format), and cryptographically-signed by an official issuing authority. However, the existing governmental and regulatory organizations are inadequate for completing such an undertaking in the necessary time frame. Perhaps, a non-profit organization without any conflicts of interests could define appropriate standards, and start an open-source ethics coding project for the rules and laws that are most urgently required by the ethical systems that we try to construct.

By standardizing ethics modules, systems from different manufacturers will use identical ethics modules that are issued by central ethics authorities. The concept of central ethics authorities might sound like part of a dystopic dictatorship, but acting ethically is mostly just a matter of obeying laws and regulations, which are a normal and necessary part of every stable society. When new laws, regulations, or bug fixes to previous modules are released, the new ethics module can be made available to all affected systems, like Microsoft Windows operating system updates; even for vehicles and robots whose manufacturer has gone out of business.

By comparison, Google’s Android operating system update mechanism is a classic example of the Law of Inevitable Ethical Inadequacy. Because Android was designed only to be effective, not ethical, hundreds of millions of Google Android devices will never receive any security patches, by design, for the simple reason that Google prioritized its profits over ethical consumer safety. We certainly don’t want self-driving vehicles, robots, and autonomous weapons systems relying on an ethics update mechanism that is as incompetently and unethically designed as Google’s Android operating system!

Such unethical and negligent corporate behaviour must be legislated out of existence, otherwise it will keep repeating itself in millions of different and damaging ways. For example, ethically inadequate Internet-of-Things devices that send unencrypted data over the internet, are vulnerable to being hacked, and will never receive security patches.

## Ethical Regulators and Super-Ethical Systems

Importing or selling such unethical devices that threaten the security of our digital infrastructure should be as illegal as selling exploding cars.

### Classification Framework

Now let's consider where the Ethical Regulator Theorem fits into the existing cybernetics framework. Table 1 lists some of the cybernetic community's definitions of first- and second-order cybernetics, as summarized by Stuart Umpleby (2001).

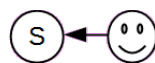
**Table 1: Definitions of first- and second-order cybernetics**

Author	First-Order Cybernetics	Second-Order Cybernetics
Von Foerster	The cybernetics of observed systems	The cybernetics of observing systems
Pask	The purpose of a model	The purpose of the modeler
Valera	Controlled systems	Autonomous systems
Umpleby	Interaction among the variables in a system	Interaction between observer and observed
Umpleby	Theories of social systems	Theories of the interaction between ideas and society

Although every one of these definitions captures an important distinction, when compared to the rigorous precision with which other scientific communities use the qualifiers “first-order” and “second-order”, the cybernetic community's use of “first-order” and “second-order” appears to be rather subjective, lacks the consensus that is required by the scientific principle, and is of little utility (Kuhn 1962).

This disarray in defining cybernetics as first-order and second-order not only prevents it from being useful to classify different types of systems, but it also prevents the classification from being extended to higher orders, which can be viewed as either a self-limiting dead-end, or paradigmatic autoapoptosis (self-programmed death), which is not entirely unlike the situation of the members of the Heaven's Gate millennial death-cult, who believed that by committing suicide, they would be rescued by an alien spacecraft and “graduate to the Next Level”.

To illustrate the problem of classifying cybernetics into “orders”, let's start by considering first-order cybernetics, which is concerned with a system, S, that is studied by an observer, as illustrated in figure 2.

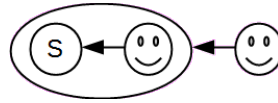


**Figure 2: First-Order Cybernetics**



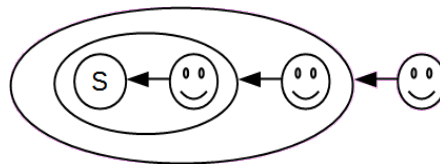
## Ethical Regulators and Super-Ethical Systems

Second-order cybernetics introduces a second observer's viewpoint, as shown in figure 3.



**Figure 3: Second-Order Cybernetics**

Logically, third-order cybernetics would add a third observer's perspective, as shown in figure 4.



**Figure 4: Third-Order Cybernetics**

However, from the perspective of the third observer, this looks more like psychology than cybernetics. In fact, the third-order cybernetics structure is isomorphic to a typical management team evaluation exercise, where the details of the task that is given to the team to work on is virtually irrelevant to the outermost observer. The innermost task could be any activity that has a goal, for example, building the highest stable tower possible out of a limited set of Lego bricks, trying to solve an impossible puzzle in a limited amount of time, or studying a first-order cybernetic system.

At even higher orders, with  $N$  observers, it becomes even more compelling that cybernetics would stop being cybernetics and become psychology.

### New Classification

It would be of more utility to define unambiguous “levels” of cybernetic systems that include categories of future systems that are already anticipated and associate each level with established concepts. To that end, the following framework for classifying cybernetic systems is proposed.

**Table 2: Framework for classifying cybernetic systems**

Level	The cybernetics of	Also known as	The cybernetician
1	Simple systems	First-order cybernetics	Watches the system
2	Complex systems	Second-order cybernetics	Participates in the system

## Ethical Regulators and Super-Ethical Systems

Level	The cybernetics of	Also known as	The cybernetician
3	Ethical systems	Cybernetics	Designs the system
4	Superintelligent systems	Technological singularity	Stares incredulously, as the system redesigns itself
5	“Super-Ethical” systems (Superintelligent and ethically adequate)	Technological utopia	Is protected by the system
6	“Super-Unethical” systems (Superintelligent and ethically inadequate)	Technological dystopia	Obeys the system

Today, in this paradigm, we are in the transition from building complex cybernetic level two systems (CL2) to building ethical/superintelligent cybernetic systems of levels three or four (CL3/4), and the future of our species and planet is in our hands. But first, a few words about each level.

### *Cybernetic Level 1: Simple Systems*

This is the domain of first-order cybernetics: Studying and designing simple systems that are effective.

### *Cybernetic Level 2: Complex Systems*

This is the domain of second-order cybernetics: Studying and designing complex systems that are effective. All observers are participants and all participants are observers. There is still much valuable and important work to be done at this level.

### *Cybernetic Level 3: Ethical Systems*

Decades ahead of his time, it was the wonderful and inspiring Ranulph Glanville (1986) who defined “cybernetics” as “the cybernetics of ethics and the ethics of cybernetics”.

The Ethical Regulator Theorem belongs at this level, which is concerned with designing man-made systems that satisfy all nine requisites of the Ethical Regulator Theorem, where the regulating agents are humans, machines, or cyberanthropic hybrids. Ethical machines must accept standardized, certified ethics modules. At this level, natural systems are of only marginal interest.

In retrospect, now that we’re not trying to extrapolate from just two data-points in concept-space, if level three is ethical systems, it’s suddenly apparent that the third observer in the third-order cybernetics system shown in Figure 4 is not necessarily a psychologist or a lost cybernetician, but could be the second observer’s conscience; her super-ego, or higher-self; that constantly self-observing sense that we all have that knows the difference between right and wrong, between good and evil, that triggers a feeling of guilt (only in non-psychopaths) if it is ignored. This self-monitoring mechanism is known as integrity, and is something that today’s ethically indifferent scientists, politicians, lawyers, bankers, and CEOs are woefully lacking.

## Ethical Regulators and Super-Ethical Systems

### *Cybernetic Level 4: Superintelligent Systems*

The technological singularity is a hypothetical moment when a self-improvement process causes runaway improvements in intelligence that results in superintelligence that is far greater than any human mind. For this to happen, the system must be sufficiently self-aware to understand its own software and/or hardware.

### **Superintelligence Tests**

These levels of self-awareness give rise to three levels of superintelligence tests. The ability to reprogram better software for itself, the ability to redesign better hardware for itself, and the ability to do both.

Together with the Turing Test (Turing 1950), these tests mark milestones in the evolution of AI systems towards superintelligence, and should cause us alarm if progress towards them is made without significant progress creating ethical systems first. Of these tests, the Turing Test is probably the easiest to achieve, because it only requires that a computer can imitate a (not necessarily very intelligent) human sufficiently well to convince humans most of the time that it is a human being, and does not require self-awareness or runaway improvements in intelligence.

### **Prophecies of Possible Futures**

In 1952, Ross Ashby wrote in his journal that super-clever machines could create a technological utopia: “It may be found that we shall solve our social problems by directing machines that can deliver an intelligence that is not our own.” (Ashby 1952a).

Two pages later, he described a technological dystopia that sounds like Google on steroids: “What people could resist propaganda and blarney directed by an I.Q. of 1,000,000? It would get to know their secret wishes, their unconscious drives; it would use symbolic messages that they didn’t understand consciously; it would play on their enthusiasms and hopes. They would be as children to it. (This sounds very much like Goebbels controlling the Germans).”.

On the appearance of such a machine, he described a paradox of perception of higher intelligence: “It seems, therefore that a super-clever machine will not look clever. It will look either deceptively simple or, more likely, merely random.” (Ashby 1952b). On the same subject, Arthur C. Clarke’s Third Law states: “Any sufficiently advanced technology is indistinguishable from magic.” (Clarke 1973). If you think that Clarke’s “magic” and Ashby’s “deceptively simple or merely random” are incompatible; take a moment to reflect on the magical simplicity and “randomness” of Google’s search results pages or an amazing Las Vegas magic show.

Just as there are two diametrically opposite archetypes for genius; namely the benevolent good genius and the nasty evil genius, it is important not to conflate systems that are ethical with ones that are not ethical by making them share the same name or category, such as “superintelligent”. To do so, would focus attention on the least important characteristic, and ignore the most important characteristic; good and evil.

## Ethical Regulators and Super-Ethical Systems

### *Cybernetic Level 5: Super-Ethical Systems*

The term “super-ethical” is proposed to refer to superintelligent systems that satisfy all nine requisites of the Ethical Regulator Theorem. Of course, by the time that super-ethical systems exist, a friendlier name will have emerged and the term “super-ethical” will seem quaintly archaic.

### *Cybernetic Level 6: Super-Unethical Systems*

The term “super-unethical” is proposed to refer to superintelligent systems that do not satisfy all nine requisites of the Ethical Regulator Theorem. This term should always carry a certain stigma, like “weapons of mass destruction”. Let no one who is working to create intelligent systems escape admitting whether the systems are, by their design or implementation, ethically inadequate.

Just as human genetic experimentation is strictly ethically regulated, we need legislation, regulation, standards, and certification to ensure that autonomous AI systems that make decisions that can have ethical consequences are subjected to the same kind of obsessively rigorous safety-oriented design, construction, and operating procedures as nuclear power stations, commercial aircraft, and vehicles that carry humans into space.

Some may start arguing that intelligence is ethically neutral, and it is, but that family of arguments are fallacies because a hyper-genius “Million I.Q. Engine” without ethics is not ethically neutral; it should be regarded like a bomb that is capable of completely destroying our planet. For a corporation even to plan to construct such a device should be a crime against humanity.

As a thought experiment, let’s imagine a hypothetical superintelligent version of Google, named the “Googlevil” corporation. The CEO is Dr. Evil, and both are without ethics, avoid transparency, and will do anything to maximize their profits and power. The corporation’s secret mission statement is “Collect and organize the world’s personal information and make it accessible and useful for maximizing our profits, power, influence, and ability to avoid paying taxes.” and its secret corporate motto is “Say ‘Don’t be evil’ then do it anyway.”

Anytime that the superintelligent Googlevil artificial intelligence or the psychopathic demagogue Dr. Evil wants to blackmail the CEOs of other corporations, politicians that can’t be bought, jury members, or Supreme Court justices around the world to make “random” decisions that incrementally further their secret mission, would they have to do anything more than query the Googlevil user-profile database? In theory, they would only need to be able to blackmail a majority of members of lower- and upper-houses to be able to get any legislation that they want in any country. Or just two or three Supreme Court justices to steer a nation into a fascist dystopia.

By the time that super-unethical AI systems exist, they will be indistinguishable from the corporations that they control. They will be immoral, immortal, enjoy legal personhood, pay no taxes, and make unlimited donations to all Googlevil-friendly political parties in all techno-democratic dystopias on the planet.

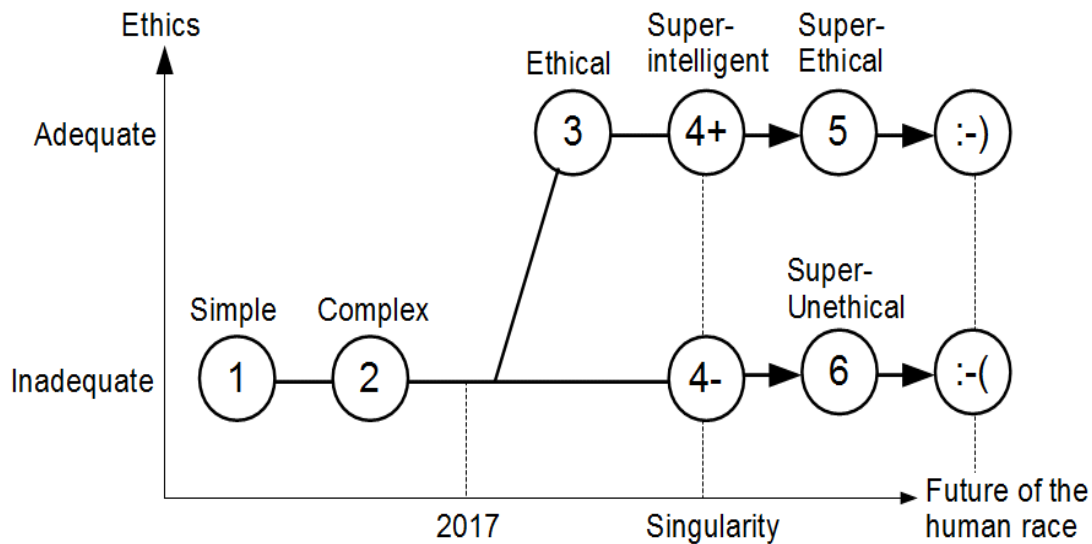
## Ethical Regulators and Super-Ethical Systems

### Future Time-Line Bifurcation Race Condition

At this point in time, there is a possibility-space bifurcation in our future time-line. Depending on whether the systems that achieve the singularity are ethically adequate or not, the runaway increase in intelligence and inevitable ethical polarization pressures will result in one of two outcomes:

- Good hyper-genius AIs protect the human race.
- Evil hyper-genius AIs dominate the human race.

Figure 5 illustrates how plotting the ethical dimension orthogonally to the intelligence dimension clarifies the dependencies between different cybernetic levels, and clearly shows that the ethically inadequate superintelligent systems of cybernetic level four-minus (CL4-) have no dependency on us first achieving ethical systems (CL3).



**Figure 5: Two mutually exclusive possible futures**

So there is a race condition that will determine which of two mutually exclusive possible futures will be the fate of our species; will our technological progress reach CL3 or CL4-first? And will legislators regulate such developments ethically and adequately, or will they sell us out to Dr. Evil's special interest lobby groups and think-tanks that will campaign vigorously for "self-regulation" — and we all know what that really means.

If we take the direct route from complex systems (CL2) to superintelligent systems that are ethically inadequate (CL4-), we quickly arrive at a dystopia that is ruled by super-unethical systems (CL6), and the utopia of being ruled by benevolent super-ethical systems (CL5) becomes permanently unreachable.

It cannot be overemphasized that CL4± is the point-of-no-return where humans could lose control over machines that become our intellectual superiors. And this is the window of opportunity to ensure that superintelligent machines are programmed with ethics and purposes that serve the greater good of humanity and the planet.

## Ethical Regulators and Super-Ethical Systems

### Universality

Anyone who has the impression that the Ethical Regulator Theorem applies primarily to artificial intelligence, self-driving vehicles, robots, and autonomous weapons systems is urged to consider how the theorem can be applied to human systems that make decisions that affect people or the environment, such as the CEO of a corporation, the U.S. Supreme Court, a political system, or yourself.

As members of a human society, we are all cybernetic regulators; of ourselves and of each other. As a thought experiment, to become a more effective and ethical force for good, you could consider ways to improve each ethical requisite as it applies to yourself, as illustrated in Table 3.

**Table 3: Ways to become a better ethical regulator**

Requisite	Example self-improvement actions
Truth	To become a good judge (effective and ethical) of who tells the truth and who distorts it, seek a wide-spectrum of opinions by finding alternative information sources that are genuinely independent of your primary sources. Investigate any inconsistencies that you notice, modify the reputation of liars appropriately, and resolve to avoid them in future.
Variety	Brainstorm new actions, responses, and strategies that you have never previously considered.
Predictability	Improve your model of human behaviour by studying the following Wikipedia articles until you are competent at recognizing the patterns in yourself and others: <ul style="list-style-type: none"> <li>• <a href="#">List of cognitive biases</a></li> <li>• <a href="#">List of fallacies</a></li> <li>• <a href="#">Defence mechanisms</a></li> <li>• <a href="#">Demagogue</a></li> </ul>
Purpose	Write down your five most important life goals: <ol style="list-style-type: none"> <li>1.</li> <li>2.</li> <li>3.</li> <li>4.</li> <li>5.</li> </ol>
Ethics	Write down five undesirable, unethical, or disrespectful behaviours that, up until now, you have tolerated in other people, organizations, or corporations: <ol style="list-style-type: none"> <li>1.</li> <li>2.</li> <li>3.</li> <li>4.</li> <li>5.</li> </ol>

## Ethical Regulators and Super-Ethical Systems

Requisite	Example self-improvement actions
	Now, next to them, write down five undesirable, unethical, or disrespectful behaviours that, up until now, you have tolerated in yourself. If you can't think of five things about yourself, read the Wikipedia article: <a href="#">Denial</a> . If that doesn't help, ask someone that you live with to suggest five things that you do that they'd prefer you not to do.
Intelligence	Read a book or take a course on personal effectiveness or critical thinking.
Influence	Identify ways that you can increase your influence (on your family, friends, colleagues, or society) to achieve your life goals and promote your ethical values.
Integrity	Seek to stop or prevent all the undesirable, unethical, or disrespectful behaviours that you listed under requisite ethics.
Transparency	Let other people know about the changes that you are making.

Finally, keep reviewing and refining your answers until they resonate with who you are and how you want your world to become.

### Our Future Epilog or Eulogy

We are approaching a decisive fork in the road in the evolution of several life forms on this planet (for example, intelligent machines, immortal corporations, political systems, and human society), and it is imperative that we learn to make these systems rigorously ethical before artificially intelligent machines reach the technological singularity, start to evolve exponentially, exceed human intelligence, and are used by ethically inadequate corporations to dominate the human race politically and economically.

For we are the generation that had the chance to steer the fate of future generations of humanity towards being ruled (potentially for eternity) by benevolent super-ethical systems that create a stable cyberanthropic utopia for us, rather than being ruled by unethical superintelligences that either enslave most of us in a cybermisanthropic dystopia (if we're lucky) or cause the extinction of our species to become a footnote in Gaia's geological record.

### Possible Future

Imagine how different the world would be:

- If we were ruled by super-ethical artificial intelligences that eliminated poverty, environmental destruction, global warming, and injustice.
- If the United Nations could deploy heavily armed peace-keeping robot armies into conflict zones to protect civilians and enforce ceasefires.

## **Ethical Regulators and Super-Ethical Systems**

- If our towns and cities were policed by robocops that protect citizens 24x7, never pepper spray, brutalize, or arrest protesters for exercising their legal right to protest peacefully, and never shoot young men in the back for being black.
- If robotic child-care-companions accompanied our children, protecting them from dangers such as physical, emotional, and sexual abuse.
- If all corporations were run ethically.

That future is possible; but only if we learn to recognize and act together in accordance with the fact: Ethics are a higher power for good that transcends science, politics, nations, and religions.

### **The Path Forwards**

To start steering the future of the human race and our wonderful planet towards becoming a stable cyberanthropic super-ethical society, I propose establishing an independent non-profit ethics institute.

#### **Research and Development**

The ethics institute will promote theoretical and practical progress:

- Coordinate and fund research into creating ethical systems and making existing systems ethical.
- Develop a taxonomy of cryptographically-signed open-source ethics modules for different types of rules, regulations, and laws that can be used by anyone, free of charge.

#### **Standards and Certification**

The ethics institute will create an ethical certification infrastructure:

- Establish standards for certifying the ethical adequacy of systems.
- Establish a curriculum for training accredited ethical consultants.
- Coordinate and regulate contracts for ethical audits and certifications.

#### **Legislation and Democracy**

The ethics institute will lobby governments to implement ethically adequate legislation and will evaluate the adequacy of any proposed legislation. In particular, promoting the following changes:

- Regulate autonomous machines to require that their design and implementation is ethically adequate, and that they support compulsory ethics modules.
- Make it illegal to import or sell products that have not been certified as being ethically adequate, unless they are excluded from requiring certification.



## Ethical Regulators and Super-Ethical Systems

- Require that all new systems and processes are designed to be ethically adequate.
- Require that all systemically important existing systems and processes are evaluated for ethical adequacy, and that any weaknesses are addressed.

### Unethical Arguments

Finally, the logical capstone to this scientific manifesto for a global ethical revolution to create a stable cyberanthropic super-ethical society is to define **The Law of Unethical Arguments**, which states: “Because no ethical argument can exist against making a system ethical, anyone who argues against this objective (or abuses its sincere supporters) is either objectively unethical, corrupt, or evil.”.

### The Final Battle

Because the human race is facing the extreme and imminent possibility of either being protected by super-ethical hyper-genius AI systems in a cyberanthropic utopia or being dominated by super-unethical hyper-genius AI systems in a cybermisanthropic dystopia, it is not irrational to view this proposed global ethical revolution as part of a final decisive battle between the forces of good and evil on this planet. Passively doing nothing makes the demagogues and psychopaths stronger. It’s time to decide which side you are on and commit to it; either you’re with us, or you’re against us.

### Spiritual Resonance

If you distil different solutions that contain alcohol, you get pure alcohol. And if you distil different religions and philosophies that contain ethics, you get pure ethics. And because ethics are a higher power for good that transcends science, politics, nations, and religions, it is probably the only force that can unify humanity to work together for our greater good.

This ethical revolution is neither a new religion nor a political movement. It is simply the product of a compassionate heart and mind, generating coherent ethical interventions in multiple complex systems (for example, but not limited to, the computational, corporate, cybernetic, personal, political, psychological, scientific, social, and spiritual realms) for the greater good of humanity, backed by the power of the Ethical Regulator Theorem, and resonating, not only with each other, but also across space and time with all good people who have ever existed — or ever will.

*In the words of Albert Einstein (1879-1955):*

- No problem can be solved from the same level of consciousness that created it.

*In the words of Mahatma Gandhi (1869-1948):*

- First they ignore you, then they laugh at you, then they fight you, then you win.
- You must become the change you wish to see in the world.
- The future depends on what you do today.

## Ethical Regulators and Super-Ethical Systems

- Happiness is when what you think, what you say, and what you do are in harmony.
- The difference between what we do and what we are capable of doing would suffice to solve most of the world's problems.
- If I have the belief that I can do it, I shall surely acquire the capacity to do it even if I may not have it at the beginning.
- Non-cooperation with evil is as much a duty as is cooperation with good.
- Capital as such is not evil; it is its wrong use that is evil.
- Poverty is the worst form of violence.
- There is sufficiency in the world for man's need, but not for man's greed.
- There are people in the world so hungry, that God cannot appear to them except in the form of bread.
- God has no religion.
- Where love is, there God is also.
- Jesus is ideal and wonderful, but you Christians — you are not like him.
- Those who say religion has nothing to do with politics do not know what religion is.
- There is a higher court than the courts of justice and that is the court of conscience.
- Victory attained by violence is tantamount to a defeat, for it is momentary.
- What difference does it make to the dead, the orphans, and the homeless, whether the mad destruction is wrought under the name of totalitarianism or the holy name of liberty or democracy?
- They may torture my body, break my bones, even kill me. Then they will have my dead body, but not my obedience.
- Your beliefs become your thoughts, your thoughts become your words, your words become your actions, your actions become your habits, your habits become your values, your values become your destiny.

*In the words of His Holiness the Dalai Lama XIV:*

- Irrespective of whether we are believers or agnostics, whether we believe in God or karma, moral ethics is a code which everyone is able to pursue.
- The ultimate authority must always rest with the individual's own reason and critical analysis.
- The true hero is one who conquers his own anger and hatred.

## **Ethical Regulators and Super-Ethical Systems**

- A good friend who points out mistakes and imperfections and rebukes evil is to be respected as if he reveals the secret of some hidden treasure.
- A lack of transparency results in distrust and a deep sense of insecurity.
- In our struggle for freedom, truth is the only weapon we possess.
- Where ignorance is our master, there is no possibility of real peace.
- Through violence, you may “solve” one problem, but you sow the seeds for another.
- I defeat my enemies by making them my friends.
- A truly compassionate attitude toward others does not change even if they behave negatively or hurt you.
- When you practice gratefulness, there is a sense of respect toward others.
- If you can, help others; if you cannot do that, at least do not harm them.
- The purpose of all the major religious traditions is not to construct big temples on the outside, but to create temples of goodness and compassion inside our hearts.
- The whole purpose of religion is to facilitate love and compassion, patience, tolerance, humility, and forgiveness.
- Love is the absence of judgement.
- Love and compassion are necessities, not luxuries. Without them humanity cannot survive.
- The more you are motivated by love, the more fearless and free your action will be.
- Be kind when possible. It is always possible.
- Don't ever mistake my silence for ignorance, my calmness for acceptance or my kindness for weakness. Compassion and tolerance are not a sign of weakness, but a sign of strength.
- The ultimate source of happiness is not money and power, but warm-heartedness.
- As people alive today, we must consider future generations: a clean environment is a human right like any other. It is therefore part of our responsibility toward others to ensure that the world we pass on is as healthy, if not healthier, than we found it.
- With realization of one's own potential and self-confidence in one's abilities, one can build a better world.
- If you think you are too small to make a difference, try sleeping with a mosquito.

## Ethical Regulators and Super-Ethical Systems

*In the words of Dr. Martin Luther King Jr. (1929-1968):*

- Those who love peace must learn to organize as effectively as those who love war.
- True peace is not merely the absence of tension. It is the presence of justice.
- Injustice anywhere is a threat to justice everywhere.
- What affects one directly, affects all indirectly.
- The time is always right to do the right thing.
- You are not only responsible for what you say, but also for what you do not say.
- Every man must decide whether to walk in the light of creative altruism or in the darkness of destructive selfishness.
- We must learn that passively to accept an unjust system is to cooperate with that system, and thereby to become a participant in its evil.
- Our scientific power has outrun our spiritual power. We have guided missiles and misguided men.
- A nation that continues year after year to spend more money on military defence than on programs of social uplift is approaching spiritual doom.
- We should never forget that everything Adolf Hitler did in Germany was “legal” and everything the Hungarian freedom fighters did in Hungary was “illegal”.
- Nonviolence is directed against forces of evil rather than against persons who happen to be doing evil. It is evil that the nonviolent resister seeks to defeat, not the persons victimized by evil.
- Nonviolence means avoiding not only external physical violence but also internal violence of spirit. You not only refuse to shoot a man, but you refuse to hate him.

*In the words of His Holiness Pope Francis:*

- We all have the duty to do good.
- Everyone has his own idea of good and evil and must choose to follow the good and fight evil as he conceives them. That would be enough to make the world a better place.
- Human rights are not only violated by terrorism, repression, or assassination, but also by unfair economic structures that create huge inequalities.
- The worship of the golden calf of old has found a new and heartless image in the cult of money and the dictatorship of an economy which is faceless and lacking any truly human goal.
- Men and women are sacrificed to the idols of profit and consumption: It is the “culture of waste”. If a computer breaks it is a tragedy, but poverty, the needs and dramas of so many people end up being considered normal.

## Ethical Regulators and Super-Ethical Systems

- We must restore hope to young people, help the old, be open to the future, spread love. Be poor among the poor. We need to include the excluded and preach peace.
- Women in the church are more important than bishops and priests.
- All that is good, all that is true, all that is beautiful, God is the truth.
- Hatred is not to be carried in the name of God. War is not to be waged in the name of God!

*In the words of Nelson Mandela (1918-2013):*

- Freedom can never be taken for granted. Each generation must safeguard it and extend it. Your parents and elders sacrificed much so that you should have freedom without suffering what they did. Use this precious right to ensure that the darkness of the past never returns.
- Fools multiply when wise men are silent.
- Education is the most powerful weapon which you can use to change the world.
- It is in your hands to create a better world for all who live in it.
- Like slavery and apartheid, poverty is not natural. It is man-made and it can be overcome and eradicated by the actions of human beings.
- Overcoming poverty is not a gesture of charity. It is an act of justice.
- As long as poverty, injustice and gross inequality persist in our world, none of us can truly rest.

*In the words of Margaret Mead (1901-1978):*

- Never doubt that a small group of committed people can change the world. Indeed it is the only thing that ever has.

*In the words of Sun Tzu (500 B.C.):*

- Great results can be achieved with small forces.
- In the midst of chaos, there is also opportunity.
- There is no instance of a nation benefiting from prolonged warfare.
- The opportunity of defeating the enemy is provided by the enemy himself.
- Supreme excellence consists of breaking the enemy's resistance without fighting.
- Victorious warriors win first, then go to war, while defeated warriors go to war first and then seek to win.
- All men can see the tactics whereby I conquer, but what none can see is the strategy out of which victory is evolved.

## Ethical Regulators and Super-Ethical Systems

- Be extremely subtle even to the point of formlessness. Be extremely mysterious even to the point of soundlessness. Thereby you can be the director of the opponent's fate.
- He wins his battles by making no mistakes. Making no mistakes is what establishes the certainty of victory, for it means conquering an enemy that is already defeated.

*In the words of Bertolt Brecht (1898-1956):*

- Change the world, she needs it. (Ändere die Welt, sie braucht es.)

*In the words of Leonardo da Vinci (1452-1519):*

- I have been impressed with the urgency of doing. Knowing is not enough; we must apply. Being willing is not enough; we must do.

*In the words of Percy Bysshe Shelly (1792-1822):*

Rise, lions after the slumber  
In unvanquishable number!  
Shake your chains to earth like dew  
Which in sleep had fallen on you:  
Ye are many — they are few!

*In the words of Han Solo (1977):*

- May the Force be with you.<sup>1</sup>

Although the authors of these quotes are separated by space, time, and their affiliations, it is easy to imagine that they all share the same human ethical belief system, and that they would have no significant arguments with each other if they were all to meet in one room to plan an ethical revolution to make the world a better place.

### Conclusion

Though this paper covers many topics, these are but means; the end has been throughout to make clear what principles must be followed when one attempts to restore ethical function to a sick organism that is, as a human society, of fearful complexity. It is my faith that the new understanding may lead to better (more effective and more ethical) systems, for the need is great.

---

<sup>1</sup> We need a sense of humour too because laughter makes us stronger. In real life, like in the movies, the good guys always have a better sense of humour than the psychopaths!

## Ethical Regulators and Super-Ethical Systems

### References

- Ashby, W. Ross (1952a). Power and I.Q. have many similar properties, [www.rossashby.info/journal](http://www.rossashby.info/journal) 16:4276-4278.
- Ashby, W. Ross (1952b). Appearance of a super-clever machine, [www.rossashby.info/journal](http://www.rossashby.info/journal) 16:4279-4280.
- Ashby, W. Ross (1956). *An Introduction to Cybernetics*, Chapman and Hall, London.
- Asimov, Isaac (1942). *Handbook of Robotics, 56th Edition (2058 A.D.)*.
- Clarke, Arthur C. (1973). Hazards of Prophecy: The Failure of Imagination, in *Profiles of the Future: An Inquiry into the Limits of the Future*.
- Conant, Roger C., and Ashby, W. Ross (1970). Every good regulator of a system must be a model of that system, *Int. J. Systems Sci.* 1(2):89-97.
- Glanville, Ranulph (1986). The Cybernetics of Ethics and the Ethics of Cybernetics, *Tutorial at 21st American Society for Cybernetics Conference*, Virginia Beach, USA.
- Kuhn, Thomas (1962). *The Structure of Scientific Revolutions*, University of Chicago Press.
- Solo, Han (1977). *Star Wars Episode IV: A New Hope*, [www.youtube.com/watch?v=VH83SsL\\_fIQ](http://www.youtube.com/watch?v=VH83SsL_fIQ)
- Stevens, John Paul (2010). Opinion of Stevens J., Supreme Court of the United States. Citizens United, Appellant v. Federal Election Commission, *Legal Information Institute*, Cornell University Law School, [www.law.cornell.edu/supct/html/08-205.ZX.html](http://www.law.cornell.edu/supct/html/08-205.ZX.html)
- Sun Tzu (500 B.C.). *The Art of War*.
- Turing, Alan M. (1950). Computing Machinery and Intelligence, *Mind* 59:433-460.
- Umpleby, Stuart A. (2001). What comes after second order cybernetics, *Cybernetics and Human Knowing* 8(3):87-89.