Article

Ethical Regulators and Super-Ethical Systems

Mick Ashby [1,2*]

[1] Archivist of the W. Ross Ashby Digital Archive, U.K.;
[2] Trustee of the American Society for Cybernetics, U.S.A.;
* Correspondence: ethics@ashby.de;

Abstract: This paper combines the Good Regulator Theorem 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 and improving 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, cyberanthropic hybrids, organizations, corporations, or government institutions. The theorem is then used to define an ethical design process that has potentially far-reaching implications for society. A six-level framework is proposed for classifying cybernetic and superintelligent systems, which highlights the existence of a possibility-space bifurcation in our future time-line. The implementation of "super-ethical" systems is identified as an urgent imperative for humanity to avoid the danger that superintelligent machines might lead to a technological dystopia. Third-order cybernetics is defined as the cybernetics of ethical systems. Concrete actions, a grand challenge, and a vision of a super-ethical society are proposed to help steer the future of the human race and our wonderful planet towards a realistically achievable minimum viable cyberanthropic utopia.

Keywords: Ethics; Regulator; Superintelligence; Super-Ethical Systems; Requisite Variety; Cybernetics; Third-Order Cybernetics; Cyberanthropic Utopia; Grand Challenge

1. Introduction

The goal of this research is to develop a theoretical basis and a systematic process for designing systems that behave ethically.

The human race has become very good at designing systems that are effective, but we are very bad at designing systems that are reliably ethical. The majority of our social and computer-based systems are ethically fragile, lacking resilience under non-ideal conditions, and are generally vulnerable to abuse and manipulation. But we are now on the cusp of a technological wave that will thrust autonomous vehicles, robots, and other artificial intelligence (AI) systems into our daily lives, for good or bad; there will be no stopping them. And despite widespread recognition of the potential risks of creating superintelligence [1] and the need to make AI and social systems ethical, cybernetics, systems theory, and AI have no systematic process for even trying to create systems that behave ethically. Instead, we have to rely on the ad hoc skills of an ethically-motivated designer to somehow specify a system that is hopefully ethical despite the constant pressure from corporate executives to do things cheaper and faster. This is not a satisfactory solution to a problem that so urgently needs to be solved. In the context of cybernetics, this could be referred to as "The Ethics Problem".

Many people think that all technologies can be used for good or evil, but this is not true. If we consider a system like that of public health inspections of restaurants, where an inspector performs a well-structured system of evaluations in defined dimensions, such as kitchen hygiene, food storage, waste management, and signs of vermin, to identify any inadequacies and specify necessary improvements to achieve certification of hygienic adequacy; such a system can only help to make restaurants more hygienic.

Might it be possible to adapt this certification model from the public health domain to create a system that can be used to certify whether a given system is ethically adequate or inadequate? And might such a system be a solution to "The Ethics Problem"?

In this paper, the terms "ethics" and "ethical" are used in a concrete applied sense of acceptable behavior. Treating ethics as a higher human quality or as something that might be learned by neural networks is rejected.

All societies regulate the behaviour of their members by defining what behavior is acceptable or unacceptable. It is primarily through the rule of law that a society can be made safe, civilized, and ethical. And the only way that society or an individual can know or prove that something non-trivially unethical has occurred is because some kind of rule has been violated. So being pragmatic, if it's unethical to break laws, regulations, and rules, then those laws, regulations, and rules define our ethics, which is why we bother to constantly refine and try to improve them. Not all rules are defined formally in writing, some are unwritten conventions, yet in every culture, it is unacceptable to break such laws, regulations, rules, or customs.

But the act of deciding what is ethical behaviour is very different to the act of behaving ethically by obeying a society's laws and rules. The lawmakers make ethical decisions about what behavior is acceptable in a society and which is forbidden, but a law-abiding citizen (or machine) needs only to obey the appropriate laws and rules in order to behave safely and ethically in most situations, with an acceptably small risk that something dangerous or unethical might result despite following the laws and rules.

And just as a law-abiding citizen does not need to be involved in the ethical decisions that are required when making laws, this paper does not address the issue of how society decides what behavior is ethical. The paper is concerned rather with how to create effective systems that are certifiably law-abiding.

None of us are ever likely to have to decide whether to switch a runaway train to a different track to reduce the number of fatalities, but if a society decides, for example, that in such a situation, minimizing fatalities is the ethical and legal obligation, then it becomes trivial to encode it in a law, regulation, or rule so that it can be understood and obeyed by humans and machines. By doing so, what was an ethical dilemma is reduced to a simple rule. This line of reasoning implies that it is sufficient to disambiguate our laws and make robots, artificial intelligence, and autonomous vehicles rigorously law abiding. It is suggested that there is absolutely no need to make such autonomous systems capable of resolving genuine ethical dilemmas, which is the job of society's lawmakers and regulatory organizations to anticipate, resolve, and codify in advance.

1.1 Literature

The starting point for this research was trying to find answers to the following question: "What characteristics must a system have for it to behave ethically?"

The existing cybernetics literature provided the first two characteristics. Conant and Ashby's Good Regulator Theorem [2] proved that every good regulator of a system must be a model of that system, but it does not specify *how* to create a good regulator. And Ashby's Law of Requisite Variety [3] 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. An ethical system must have more than just these two characteristics.

Recent approaches to making artificial intelligence ethical, such as IBM's "Everyday Ethics for Artificial Intelligence: A practical guide for designers and developers" [4] and the European Commission's "High-Level Expert Group on Artificial Intelligence: Draft Ethics Guidelines for Trustworthy AI" [5], merely provide a wish list of requirements without offering anything that can be applied systematically to design an ethical AI.

Heinz von Foerster proposed an Ethical Imperative: "Act always so as to increase the number of choices" [6]. Although this principle is valuable in the context of psychological therapy, it specifies no end condition, i.e. when to stop adding more choices. If one were to apply it when deciding how many different types of propulsion systems to build into a manned spacecraft to adjust its motion

 doi:10.20944/preprints201907.0349.v1

and orientation, it would lead to unnecessary choices, extra costs, extra weight, increase the number of points of possible failure, and therefore increase the risk of catastrophic failure and loss of life. This counter example shows that maximizing choice can be the wrong (unethical) thing to do. And by definition, implementing more choices than is necessary to achieve the goal of a system is unnecessary. So, we must reject von Foerster's Ethical Imperative as being flawed. □

In 1990, von Foerster gave a lecture titled "Ethics and Second-Order Cybernetics" to the International Conference, Systems and Family Therapy: Ethics, Epistemology, New Methods, in Paris, France [7]. However, despite its promising title, it provides nothing concrete or systematic for making systems ethical.

Stafford Beer's viable system model [8] is specific to hierarchically structured systems and associates ethics with a specific level of the hierarchy (System 5). But rather like creating an ethics committee, assigning "ethics" to a level of an architecture is insufficient to make a system ethical, it does not explain *how* to make the system ethical. It just creates the illusion of having solved the problem, but the problem has not been solved; only delegated. By contrast, we expect reliable ethicalness to be an inevitable emergent property of the entire system — if and only if the system is ethically adequate.

1.2 Methodology

An important early step was to realize that the Good Regulator Theorem 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 meaning, and only uses "good" when both meanings are intended. It is only by imposing precision in the use of terminology that it was possible to clarify the otherwise muddled thinking and isolate the essence of an ethical system.

To identify more necessary characteristics, a selection of ethical and unethical systems were subjected to analysis, including an autonomous vehicle, a bank ATM, capitalism, a central bank, a corrupt politician, a dating system, democracy, a healthcare robot, a jury, a law-abiding citizen, a money laundering bank, a product design process, a superintelligent machine, the U.S. Supreme Court system, a vehicle exhaust emission test cheating corporation, and a voting machine. Considering these 16 diverse systems helped identify more characteristics, such as having ethical goals, laws, and the intelligence to understand the laws and make rational decisions.

Some other necessary characteristics only became apparent after looking for ways that an evil actor (internal or external to the system) could subvert each system, such as by hacking, tampering, feeding the system with false information, or by threatening, bribing and blackmailing people who have influence on the system. Then a minimum set of additional characteristics were sought that would counter all of the identified potential vulnerabilities.

In all, nine characteristics were identified that are necessary and sufficient for a system to behave ethically. These nine requisites are integrated in the Ethical Regulator Theorem (ERT), which can be used as a decision function, IsEthical, that can be applied systematically to categorize any system as being ethically adequate, ethically inadequate, or ethically undecidable. A proof of the theorem is provided. Another result of ERT is a basis (known as the MakeEthical function) for systematically identifying improvements that are necessary for a given system to be made ethically adequate. The IsEthical and MakeEthical functions can be used to construct an ethical design process.

Because ERT did not seem to fit in the existing cybernetics framework, a new framework was developed out of necessity. It uses the IsEthical function to distinguish between two types of superintelligent machines; those that are ethically adequate and those that are ethically inadequate. Together, the superintelligence and ethics dimensions are used to identify four well-defined classes of systems. These four distinct classes can be appended to the existing two levels of first-order and second-order cybernetic systems to create a six-level framework for classifying cybernetic and superintelligent systems. An unexpected consequence of trying to categorize ERT was the realization that third-order cybernetics should be defined as "the cybernetics of ethical systems".

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

Because the Ethical Regulator Theorem can be applied to any system and offers a new and systematic approach to making systems more ethical, the implications for making the world a better place are significant and should be explored further.

One result of the exploration of the proposed six-level framework is the identification of a race condition that results in either a cyberanthropic utopia or a cybermisanthropic dystopia. This dystopic threat is well known, however, by identifying the exact nature of the race condition, it becomes clear what strategy must be employed to try to avoid the possibility that superintelligent machines could lead to a dystopian disaster.

Because it is imperative for humanity to avoid this existential threat, concrete actions are proposed, including a grand challenge to apply ERT to new and existing systems in all areas of society in what is characterized as a systemic ethical revolution. And because a key component of that revolution is psychological, 80 ethically inspiring quotes from 10 famous people from five continents are presented that demonstrate that ethics transcends science, politics, nations, and religions, and is probably the only force that can unify humanity to work together for our greater good.

2 The Ethical Regulator Theorem

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

1. **Purpose** expressed as unambiguously prioritized goals.
2. **Truth** about the past and present.
3. **Variety** of possible actions.
4. **Predictability** of the future effects of actions.
5. **Intelligence** to choose the best actions.
6. **Influence** on the system being regulated.
7. **Ethics** expressed as unambiguously prioritized rules.
8. **Integrity** of all subsystems.
9. **Transparency** of ethical behavior.

Of these nine requisites, only the first six are necessary for a regulator to be effective. If a system does not need to be ethical, the three requisites ethics, integrity, and transparency are optional. Figure 1 and the following sections explain the requisites in more detail.

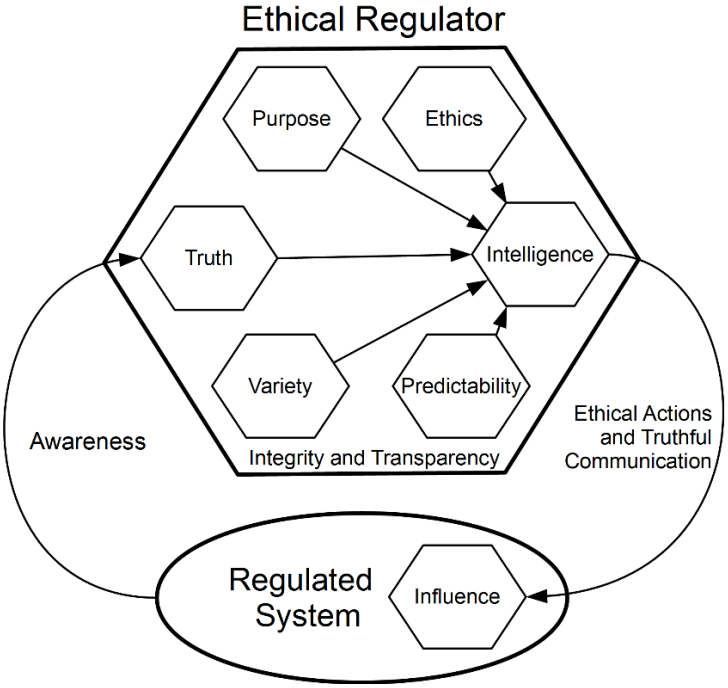

Figure 1: The Ethical Regulator System

 doi:10.20944/preprints201907.0349.v1

2.1 Requisite Purpose

Because complex systems are required to satisfy multiple goals, **Purpose** must be expressed as unambiguously prioritized goals. Without well-defined goals, the system cannot be effective and might randomly adopt or default to a goal that is unethical.

2.2 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, knowledge, and beliefs. 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.

An ethical regulator doesn't require perfectly accurate information, but it must be sufficiently truth-seeking to be able to cope 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.

2.3 Requisite Variety

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

2.4 Requisite Predictability

Predictability requires a sufficiently accurate model of the regulator and the system being regulated, to be able to rank the actions and strategies that will give the best outcome. This is nothing other than the Good Regulator Theorem.

2.5 Requisite Intelligence

Intelligence is applied to the previous requisite types of information to select the most effective/ethical strategies and actions from the set of possible actions. And because the output of one regulator is generally an input to other regulators (systems or people), if the selected action is an act of communication, it must be as truthful as possible.

2.6 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, but a function of the connectivity relationships that span from the regulator's outputs to elements of the regulated system and its environment. If a regulator has no influence on the regulated system, it isn't a true regulator, it is an observer or simulation, and there are no direct ethical consequences; which can be important when observing or simulating dangerous situations.

Depending on the nature of the system that is being regulated, the speed and duration of the effects of actions can vary greatly. For example, a self-driving vehicle applying the brakes has a brief yet immediate effect; the effects of the Supreme Court issuing a ruling are much slower but could last for decades or possibly centuries; and the cascade caused by someone sending a message to a complex network of amplifying/attenuating variable-delay transmission repeaters, known as Twitter followers, is unpredictably chaotic.

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.

 doi:10.20944/preprints201907.0349.v1

Let us consider two systems, A and B, that 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 effectiveness, the use of a gun does not amplify his variety, it is just one existing element in his range of possible 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.

In the light of the concept of influence, the belief that variety can be amplified appears to be as delusional as the idea that randomness can be amplified. Feeding variety or randomness into a genuinely noiseless amplifier cannot produce more variety or randomness than was fed into it. The variety of the robber or an advertising message is effectively constant.

The six requisites described so far are necessary and sufficient for a system to be effective but are not sufficient for it to be ethical.

2.7 Effectiveness Function

The Ethical Regulator Theorem implies that we can define a function for the effectiveness that a regulator, R, has in controlling a system. It captures how the effectiveness of the regulator depends on the effectiveness of all six requisites:

$$\text{Effectiveness}_R = \text{Purpose}_R \times \text{Truth}_R \times \text{Variety}_R \times \text{Predictability}_R \times \text{Intelligence}_R \times \text{Influence}_R \quad (1)$$

In this form, we would assign each requisite an effectiveness value between 0 and 1, where 1 means that it is perfect or optimal. And if the effectiveness of even just one of the requisites is close to zero, the effectiveness of the whole regulator is massively reduced. Applied to our two politicians: If $\text{Effectiveness}_A > \text{Effectiveness}_B$, then A is more likely than B to win control of system C.

However, it is neither necessary nor possible to calculate meaningful numerical values 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 understanding of the effectiveness function informs the system design strategy; recognizing that a maximally effective system requires that the effectiveness of these six requisite dimensions are maximized, and that a successful attack on the integrity or effectiveness of any of them spells disaster for the effectiveness of the whole system.

It is worth noting that in social systems, money can buy media influence; and if the media is broadcasting lies, propaganda, or advertising, it reduces the quality of Truth_X that is received by every voter or consumer, X, which can manipulate them into making decisions that are not in their best interest.

2.8 Requisite Ethics

Ethics must be expressed as unambiguously prioritized laws, regulations, and rules that codify constraints and imperatives, for example, Isaac Asimov's First Law of Robotics: "A robot may not injure a human being or, through inaction, allow a human being to come to harm." [9], but ideally, expressed unambiguously in a formal language such as XML, which can be understood by humans and computers.

Ethical rules define constraints on the variety of actions and have a higher priority than the goals for purpose. By always obeying the relevant highest priority rules, the regulator is guaranteed to act ethically within the scope of the ethical schema, which provides a model of acceptable (ethical) behavior. The ethical rules have the power of veto over possible actions, which makes it safe for AI to generate candidate actions algorithmically, without having to worry whether it might generate unethical possibilities.

Because ethical schemas vary between different cultures, in machines, they must be handled as plug-ins. And because an ethical schema can encode any ethics, good or evil, each ethical schema must be anchored explicitly in the laws of a legislative jurisdiction. When a person or system crosses a state or national border it is necessary to activate a different set of ethical schemas, i.e. a different

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

set of laws, regulations, and rules. And the ethics modules must be prioritized so that it is unambiguous which module has precedence in the event of a conflict, for example, between national and state laws. The highest-level laws could be encoded in hardware to be unhackable.

A taxonomy of ethics modules can provide rules for all conceivable situations. For example, child-care, traffic-rules, gun-law, tax-law, contract-law, maritime-law, drone-flying, police-regulations, and warfare-rules-of-engagement.

Ethics modules can be treated like device drivers, so that to be fully operational, a hypothetical gun-carrying robot that can drive on roads requires valid ethics modules for gun-law and traffic-rules. Without both ethics modules for the appropriate legal jurisdiction, the robot's gun or driving capabilities are automatically disabled.

By legislating that all autonomous artificial intelligence systems must 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; 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.

When we introduce ethics, the effectiveness function 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 something like the following:

$$\text{Effectiveness}_A = \text{Purpose}_A \times \text{Truth}_A \times (\text{Variety}_A - \text{Ethics}_A) \times \text{Predictability}_A \times \text{Intelligence}_A \times \text{Influence}_A \quad (2)$$

$$\text{Effectiveness}_B = \text{Purpose}_B \times \text{Truth}_B \times \text{Variety}_B \times \text{Predictability}_B \times \text{Intelligence}_B \times \text{Influence}_B \quad (3)$$

Which captures the reality that politicians and businessmen who lie and cheat have an advantage over ones that are ethical.

2.9 Requisite Integrity

Integrity of the regulator and its subsystems must be assured through features such as resistance to tampering, intrusion detection, cryptographically authenticated ethics modules, and compliance with all laws, regulations and rules. Monitoring mechanisms must detect if any invalid ethics modules are being used or if an ethical constraint is violated, and if necessary, activate an ethical fail-safe mode, preserve evidence, and notify the manufacturer and/or the appropriate authorities.

The regulator's first-order integrity mechanisms offer no protection to the pathways on which the regulator depends to influence the system. This poses a potential vulnerability that can only be mitigated by using the awareness feedback to check for evidence of the effect of each action.

2.10 Requisite Transparency

Demanding to be trusted is unethical because it enables betrayal. Trustworthiness must always be provable through **Transparency**. So, **The Law of Ethical Transparency** is introduced, stating:

> "**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 doesn't really matter whether the programmers of a chess playing robot can find out why a piece was sacrificed, the logic of ethical decisions must never be hidden in the depths of opaque processes, neural networks, or lost to the passage of time. Generally, this requisite can be satisfied by keeping an audit trail that is adequate and secure.

When an ethically adequate system violates an ethical constraint, as they sometimes will, analysis of the audit trail will identify the reason. For example, because a faulty neural network wrongly identified a boy leading a cow as a calf leading a man, or it will prove who knew what about illegal corporate activities.

Integrity and **Transparency** are codependent security requisites: We require both integrity of transparency and transparency of integrity.

2.11 Evaluating Ethical Adequacy

Like a public health inspection of a restaurant, an evaluated system is judged on the adequacy of each requisite dimension. If and only if a system completely satisfies all nine ERT requisites is it said to be "**ethically adequate**". Otherwise it is classified as "**ethically inadequate**" and the weaknesses listed with recommendations for improving them.

Because a truly ethical system must be maximally tamper-resistant and unhackable, the evaluation of ethical adequacy also has similarities to how a Red Team performs network penetration testing; where the evaluation team tries to identify weaknesses and theoretical possibilities to subvert the integrity of the system and all its subsystems.

For each of the nine dimensions, D_i, the evaluators must consider the following three questions:

- Is the system adequate in D_i?
- Can the system be improved in D_i?
- Can the system be subverted in D_i?

This requires that the system's adequacy is considered in 27 different ways, which delivers a thorough and systematic evaluation of the system's strengths and weaknesses.

The theorem cannot be used to certify that an ethical schema is ethical because schemas (i.e. laws, regulations, and rules) can vary arbitrarily between different cultures. However, it can be used to help identify the root causes of crises and to evaluate the ethical adequacy of any proposed interventions [10]. In the near future, accredited ethical consultants may specialize in auditing and certifying the ethical adequacy of existing and proposed, products, processes. laws, organizations, and systems.

3 Ethical Regulator Theorem Proof and Consequences

Now that we understand the nine requisites better, is it possible to prove that they are indeed necessary and sufficient for a cybernetic regulator to be effective and ethical?

3.1 Proof of Necessity

Proving necessity is simple: One-by-one, for each of the nine requisites dimensions, D_i, ask yourself the question "Can a regulator be effective or ethical without requisite D_i?" — If it can't, then D_i is necessary. For example, "Can a regulator be effective or ethical without Truth?"

The answer in each case is rather obvious, especially if you refer to Figure 1 and, one-by-one, cover each requisite using your thumb, and then consider whether the resulting system can be effective or ethical without the obscured requisite. Table 1 summarizes the results, which confirm the necessity claims, including the claim that ethics, integrity, and transparency are optional for systems that only need to be effective.

Table 1. Proof of necessity "by thumb"

Requisite Dimension	Necessary to be effective[1]?	Necessary to be ethical?
Purpose	Yes	Yes
Truth	Yes	Yes
Variety	Yes	Yes
Predictability	Yes	Yes
Intelligence	Yes	Yes
Influence	Yes	Yes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

Ethics	No	Yes
Integrity	No	Yes
Transparency	No	Yes

[1] For effectiveness, the positive results for necessity correspond to the solutions for Effectiveness$_R$ = 0. I.e. when Purpose$_R$ x Truth$_R$ x Variety$_R$ x Predictability$_R$ x Intelligence$_R$ x Influence$_R$ = 0, for example, when Truth$_R$ = 0, but not when Transparency$_R$ = 0. This agreement between the ERT effectiveness function and Table 1 is unremarkable because the effectiveness function was constructed from the results of posing the necessity question for each requisite. So, the agreement does not confirm the correctness of the theorem, but by performing this exercise yourself, you can confirm the correctness of the effectiveness function. □

3.2 Proof of Sufficiency

Proving that the nine requisites are sufficient, is not so simple. First, let us assert that in the real world, effective systems and ethical systems exist. Now, for all those such systems, do any of them rely on any information, ability, or other factor to achieve effectiveness or ethicalness that is not covered by the nine requisites?

It is claimed that for all such systems that have been considered, the answer is no. However, this claim is easily refutable because it will only take one person to find one example of a necessary factor that is not covered by the nine requisites to demolish the claim of sufficiency. In the event of that happening, we would adapt the theorem, if necessary adding another requisite, reassert the sufficiency claim, thank whoever found the missing requisite, and issue the challenge: "Okay, *now* find one!"

So, although it is impossible to prove that such an exception does not exist, we can assert that it will always be possible to extend the theorem to include any missing requisites that might be identified in the future, thus restoring the validity of the claim of sufficiency for all known systems that have been considered. □

3.3 ERT Universality

Anyone who has the impression that ERT primarily applies to artificial intelligence, robots, self-driving vehicles, 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 organizations, corporations, education systems, electoral systems, government institutions, CEOs, or yourself.

Justice Stevens [11] provides an excellent example of analyzing the ethical inadequacy of the "Citizens United" ruling. And his opinion that "The Court's ruling threatens to undermine the integrity of elected institutions across the Nation." implies that there is a pressing need to evaluate the ethical adequacy of the U.S. Supreme Court system.

Because the Ethical Regulator Theorem, i.e. the IsEthical and MakeEthical functions, can be applied to any system, the nine ERT dimensions define a domain-independent abstraction layer that can be used to map from any system/regulator to any other system/regulator. This creates a vocabulary, or isomorphism, that allows practitioners in one domain to communicate meaningfully with practitioners in seemingly unrelated domains, and share insights and solutions, for example, across artificial intelligence, corporate governance, education systems, and designing consumer products. Specialists in each domain can discuss their challenges and solutions to improving purpose, truth, variety, predictability, intelligence/strategy, influence, ethics, integrity, and transparency. For example, perhaps a cloud-based secure audit trail service that was developed for one specific domain can be used to solve the transparency and integrity requirements in a completely unrelated domain.

3.4 ERT Reflexivity and Algebra

If the Ethical Regulator Theorem is genuinely universal, it can be applied to absolutely any system. In particular, it must produce meaningful results for two special cases: When we apply ERT to itself, and when we apply ERT to second-order cybernetics (2oC).

First, let us define a convenient algebra that allows us to express important assertions in this domain. We need to distinguish between: I. The act of evaluating the ethical adequacy of a system, and II. The act of determining the set of transformations or interventions that are necessary to make a system ethically adequate:

I. A function, IsEthical(S), returns the value True if system S is ethically adequate, it returns the value False if S is ethically inadequate, or it returns the value Undecidable if S is significantly inconsistent, contradictory, or opaque. The value Undecidable should be regarded as an error message rather than a type of system.
II. A function, MakeEthical(S), returns a set of transformations or interventions to make system S ethically adequate. If S is already ethically adequate, the function returns an empty set, {}.

Now we can use this ERT algebra to make some interesting and controversial claims in Table 2:

Table 2: Some ERT algebra assertions

No.	Claim	Interpretation / Justification
1	IsEthical(ERT) = True	The ERT system fulfils all nine requisites of ERT and is therefore ethically adequate. It can only be used to make systems more ethical.
2	MakeEthical(ERT) = {}	The ERT system is sufficient to be ethically adequate. Nothing else is required.
3	IsEthical(2oC) = False	Second-order cybernetics is ethically inadequate. Unlike ERT, it has no intrinsic ethics or integrity, so it can be used to make good or evil systems. It doesn't go beyond achieving effectiveness.
4	MakeEthical(2oC) = ERT	To become ethically adequate, 2oC needs the set of ERT concepts.
5	IsEthical(2oC + ERT) = True	Nothing in 2oC is incompatible with ERT.
6	2oC + ERT = 3oC	Logically, the system that is created by joining the 2oC and ERT systems would be named third-order cybernetics (3oC).
7	IsEthical(Capitalism) = False	Capitalism is ethically inadequate.
8	MakeEthical(Capitalism) = { Ethics, Integrity, Transparency }	Capitalism might be adequate in the six requisites for effectiveness, but it is obviously deficient in Ethics (laws, regulations, and rules), Integrity (compliance), and Transparency (audit trails). These must all be increased to make capitalism ethical. □

3.5 The Law of Inevitable Ethical Inadequacy

We can build on the proof of necessity to derive this new law:

"**If you don't specify that you require a secure ethical system, what you get is an insecure unethical system.**"

The reason is because when ethical adequacy is not specified as a requirement for a system design, the resulting design phase will tend to optimize for effectiveness and maximally avoid the extra costs that would be incurred by implementing the ethics, integrity, and transparency dimensions, which are optional for a system that only needs to be effective, thus guaranteeing that the resulting system is ethically inadequate and vulnerable to manipulation; by design. □

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

Note that for some systems, the term "ethical" might include aspects such as hygienic, safe, fair, honest, law-abiding, or environmentally friendly.

4 Ethical Design Process

Figure 2 shows the generic elements of a typical design process, in which an analysis phase produces a requirements artifact, which is the input to the design phase that produces a specification artifact, which is used as the input to the implementation phase, which realizes the system.

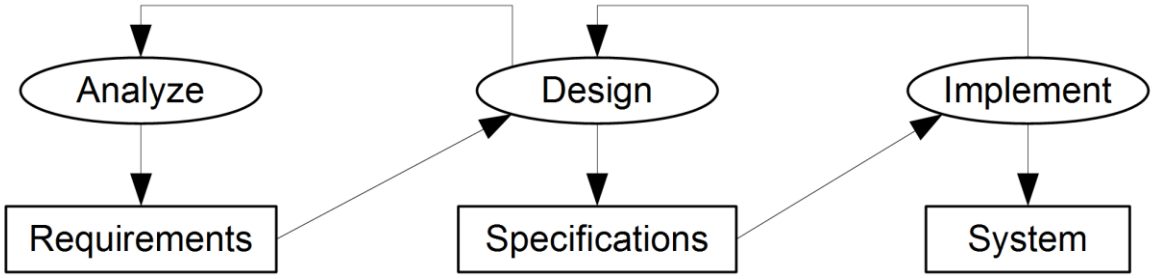

Figure 2: Ethically Inadequate Design Process

If a problem is found in the requirements during the design phase, feedback can trigger another iteration of the analysis phase. And if a problem in the specifications is found during implementation, feedback can trigger the design team to update the specifications or pass feedback to the analysis team to update the requirements.

Such a design process can be effective at producing systems that are effective, however, because the design process is ethically inadequate, it is inevitably only capable of producing systems that are also ethically inadequate; and we cannot be sure that the resulting systems are not actually ethically evil; by accident, or by design.

Fortunately, we can transform any effective but ethically inadequate design process to make it ethically adequate by simply adding ethical adequacy acceptance testing of the requirements and specifications. How we can retrofit the Ethical Regulator Theorem (ERT) to any effective design process that produces requirements and specifications before the implementation phase starts is shown in Figure 3.

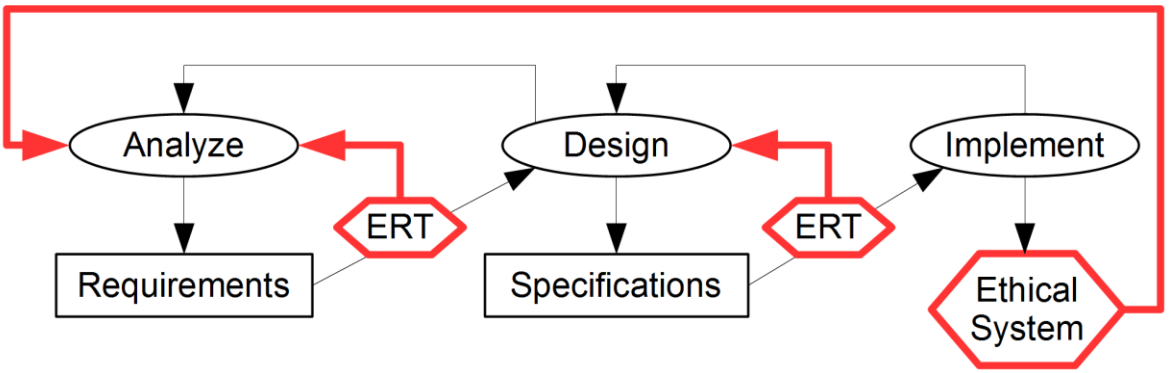

Figure 3: Ethically Adequate Design Process

Like any other quality assurance testing, the ERT IsEthical and MakeEthical evaluations should be performed by a team that was not involved in the production of the artifact being tested. If an artifact is found to be ethically inadequate, it is rejected and recommendations for fixing the problems are provided as feedback to trigger another iteration of that phase. If an artifact is found to be ethically adequate, the artifact is accepted and passed onto the next phase. Because the two ERT testing steps ensure that the requirements and specifications are ethically adequate, if the implementation process performs an effective and lossless implementation of the specifications, the resulting system will also be ethically adequate. □

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

This means that instances of the resulting system that are deployed in the real-world will include a real-time integrity monitoring mechanism that detects and report any significant problems as feedback to the analysis team, which must decide whether the problem necessitates instructing the system to activate an ethical fail-safe mode, a remote update, a reimplementation, redesigning the specifications, and/or updating the requirements. Only if the system fails to enter its ethical fail-safe mode might it be necessary for it to be deactivated using a kill switch or for it to be "retired" by a blade runner.

This concludes the description of the theorem and how to use it.

5 Discussion

The Ethical Regulator Theorem has many far-reaching implications.

5.1 Legislative Implications

By creating a well-defined interface for coding ethics, it becomes easier to apportion 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 in an emergency, decides to hit a cow to avoid hitting a dog, then the car manufacturer might be held liable for killing a sacred animal. But if the audit trail proves that the correct ethics module was activated, but the "don't hit cows" rule had an incorrectly low priority in the ethics schema, then the car manufacturer would not be liable.

It is foreseeable that one-day the laws and regulations of most countries will be published in a standardized computer-readable 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 to complete such a task 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 laws, regulations, and rules that are most urgently required by the ethically adequate systems that we try to construct.

By standardizing ethics modules, systems from different manufacturers will all use identical ethics modules that are issued by national or international 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, regulations, and rules, which are a normal and necessary part of every stable society. These ethics authorities could be independent of the legislative branch of government if the government lacks the necessary resources or commitment to unambiguous digital lawmaking.

Like Microsoft Windows operating system updates, when new laws, regulations, rules, or bug fixes to a previous ethics module are released, the new ethics module can be made available securely to all affected autonomous systems; crucially, including systems whose manufacturer has gone out of business or doesn't care about fixing end-user safety issues.

By comparison, Google's Android operating system is a classic example of the Law of Inevitable Ethical Inadequacy. Because Android was designed only to be effective, not ethical, Google delegated the responsibility for issuing Android updates to the device manufacturers. The inevitable and predictable consequence of that design decision is that most Android devices (87%) are insecure [12]. This exposes over **one billion** Android users to being hacked and their identity or credit card details stolen by criminals. The resulting chaos and the expensive suffering of the victims is not an innocent mistake, it is the direct result of Google deliberately externalizing costs onto others and prioritizing its profits over ethical consumer safety. They could have designed it differently. And if we can't trust Google, who can we trust?

We certainly don't want robots, self-driving vehicles, and autonomous weapons systems relying on an update mechanism that stops working when the manufacturer goes out of business or decides to optimize its profits at the expense of security and safety updates.

Such unethical corporate behavior must be legislated out of existence, otherwise it will keep recurring in different and damaging ways; causing unnecessary externalized costs and social chaos. 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. Importing or selling

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

5.2 Classification Framework

Now let's consider where the Ethical Regulator Theorem fits into the existing cybernetics framework. One might assume that the theorem belongs in second-order cybernetics, however, in a 1990 conference plenary presentation [7], Heinz von Foerster (who made the distinction between first- and second-order cybernetics in 1974) implied that combining ethics and second-order cybernetics is not something that he would have suggested:

"I am impressed by the ingenuity of the organizers who suggested to me the title of my presentation. They wanted me to address myself to 'Ethics and Second-Order Cybernetics'. To be honest, I would have never dared to propose such an outrageous title, but I must say that I am delighted that this title was chosen for me."

Table 3 lists some of the cybernetic community's definitions of first- and second-order cybernetics, as summarized by Stuart Umpleby [13].

Table 3. 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 system
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 how the qualifiers "first-order" and "second-order" are used by other scientific communities, the cybernetic community's use of them appears to be rather subjective, lacks the consensus that is required by the scientific principle, and is of little utility, as required by Kuhn [14].

This incoherence in defining cybernetics as first-order and second-order not only prevents it from being useful to classify different types of systems and dissipates intellectual energy, but it also prevents the classification from being extended to higher orders, which can be viewed as either a self-limiting dead-end, or paradigmal autoapoptosis (self-programmed death), which is not entirely unlike the situation of 39 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 observer-centric "orders", let's start by considering first- and second-order cybernetics, as defined by von Foerster. Figure 4 illustrates how the observers' perspectives relate to a system, S.

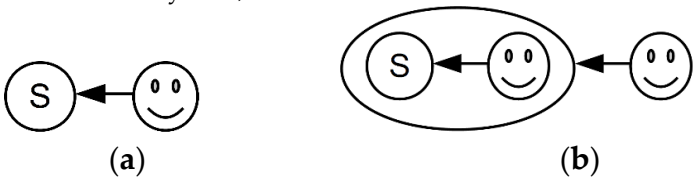

Figure 4: (**a**) First-order cybernetics (**b**) Second-order cybernetics

How can we use this paradigm to predict the future of cybernetics? Logically, third-order cybernetics would add a third observer's perspective, as shown in figure 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

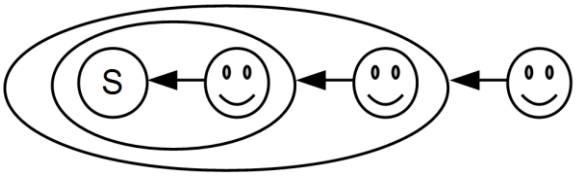

Figure 5: Third-order cybernetics

However, from the perspective of the third observer, this looks more like psychology than cybernetics. In fact, this 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. It can be any goal-oriented activity, such as building the highest stable tower possible from a limited set of Lego bricks, solving an impossible puzzle in a limited amount of time, or studying a first-order cybernetic system.

5.3 New Classification Framework

It could be of more utility to define "levels" of cybernetic systems that include categories of future systems that are already anticipated and associate each level with established concepts. To that end, Table 4 defines a six-level framework for classifying cybernetic and superintelligent systems that makes use of the ERT IsEthical function to distinguish between two important subclasses of superintelligent systems.

Table 4. Six-level framework for classifying cybernetic and superintelligent systems

Level	The cybernetics of	Also known as	The cybernetician
1	**Simple** systems	First-order cybernetics	Observes the system
2	**Complex** systems	Second-order cybernetics	Participates in the system
3	**Ethical** systems	Third-order cybernetics or Cybernethics	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 or Cyberanthropic utopia	Is protected by the system
6	**Super-Unethical** systems (Superintelligent and ethically inadequate)	Technological dystopia or Cybermisanthropic dystopia	Is manipulated to obey the system

Today, we are in the transition from building complex cybernetic level two systems (CL2) to building ethical systems and superintelligent systems of cybernetic levels three and four (CL3 and CL4), and the future of our species and our fragile ecosystem is in our hands, but first, let's clarify each level and explore where this new framework leads us.

5.3.1 Cybernetic Level 1: Simple Systems

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

5.3.2 Cybernetic Level 2: Complex Systems

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

5.3.3 Cybernetic Level 3: Ethical Systems

In 1986, decades ahead of his time, it was the wonderful and inspiring Ranulph Glanville who defined "the cybernetics of ethics and the ethics of cybernetics" as "cybernethics" [15].

The Ethical Regulator Theorem belongs at this level, which is concerned with designing man-made systems that are ethically adequate. Such systems must satisfy all nine requisites of the Ethical Regulator Theorem and the regulating agents can be humans, machines, cyberanthropic hybrids, organizations, corporations, or government institutions. Ethically adequate autonomous machines must obey certified ethics modules.

In retrospect, now that we're not trying to extrapolate from just two points in concept-space, if level three cybernetic systems are ethical, it's apparent that the third observer in the third-order cybernetics system of Figure 5 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. This self-monitoring mechanism is known as integrity, and is something that today's ethically indifferent scientists, politicians, CEOs, managers, corporations, lawyers, bankers, and billionaires are woefully lacking. In non-psychopaths, it is integrity that triggers feelings of bad conscience, regret, or guilt if it is ignored.

5.3.4 Cybernetic Level 4: Superintelligent Systems

The technological singularity is a hypothetical moment when a self-improvement process in a machine 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 of its own software and/or hardware.

5.3.5 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 [16], 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 the easiest to achieve because it is essentially a parlor game that 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.

5.3.6 Prophecies of Possible Futures

In 1951, Ross Ashby started considering how to plan an advanced society as a "super brain" [17]. A year later, he described how super-clever machines could create a cyberanthropic 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." [18]

Two pages later, he described a cybermisanthropic dystopia where a "Million I.Q. Engine" sounds like Facebook and Google, but 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)."

 doi:10.20944/preprints201907.0349.v1

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." [19]. On the same subject, Arthur C. Clarke's Third Law states: "Any sufficiently advanced technology is indistinguishable from magic." [20]. 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 a Las Vegas magic show or Google's search results' pages.

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", "Christian", or "super-rich". To do so would focus attention on the least important dimension and ignore the most important dimension: Good and Evil.

5.3.7 Cybernetic Level 5: Super-Ethical Systems

The term "super-ethical" is proposed to refer to superintelligent systems that are ethically adequate. 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.

5.3.8 Cybernetic Level 6: Super-Unethical Systems

The term "super-unethical" is proposed to refer to superintelligent systems that are ethically inadequate. This term should always carry a certain stigma, like "weapons of mass destruction". No one who is working to create artificially intelligent systems should be allowed to escape admitting whether the systems are 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 commercial aircraft, nuclear power stations, and vehicles that carry humans into space.

One could 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. Even if it had ethical goals, it might break laws to achieve them. The possibility of creating a superintelligent machine that is ethically inadequate should be treated like a bomb that could destroy our planet. Even just planning to construct such a device is effectively conspiring to commit a crime against humanity.

As a thought experiment, let's imagine a hypothetical super-unethical version of Google, named the Googlevil Corporation. The CEO is Dr. Evil, and both the CEO and the corporate AI 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, influence, and ability to avoid paying taxes." and its secret corporate mantra is "Sincerely say 'Believe me, we don't do evil', do it anyway, then look people in the eye and give them a Zuckerberg-smile!"

Anytime that the super-unethical 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 (how hard can that be?) to be able to get *any* legislation that they want in *any* country. Or just a few Supreme Court justices to steer a nation into a fascist dystopia.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

5.4 Future Time-Line Bifurcation Race Condition

At this point in time, there is an existentially critical fork 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 super-ethical AIs protect humanity.
- Evil super-unethical AIs dominate humanity.

Figure 6 illustrates how plotting the ethical dimension orthogonally to the intelligence dimension clarifies the non-linear 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 succeeding creating the ethical systems of cybernetic level three (CL3).

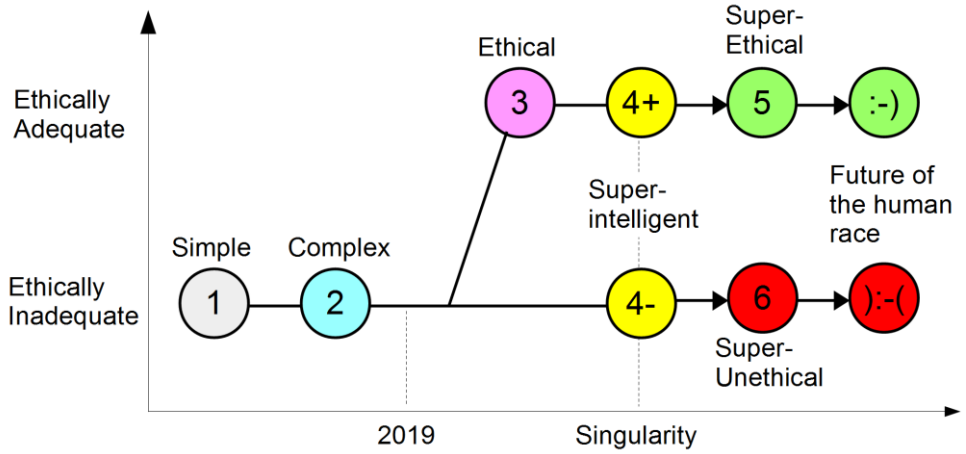

Figure 6: Two mutually exclusive possible futures

If we continue on the current path from complex systems (CL2) to ethically inadequate superintelligent systems (CL4-), we will quickly arrive in a dystopia that is dominated by super-unethical systems (CL6), and the potential cyberanthropic utopia of being ruled by benevolent super-ethical systems (CL5) will become permanently unreachable.

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

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 our fragile ecosystem. Put simply: We must create ethical systems *before* we create superintelligent systems!

In this context, it is clear that the ultimate grand challenge for cybernethics and third-order cyberneticians is to find ways to build ethical and super-ethical systems, avoid a cybermisanthropic dystopia, and help humanity create a super-ethical society.

5.5 Super-Ethical Society

Imagine how different the world would be:

- If we were happy to be ruled by benevolent super-ethical artificial intelligences that eliminated poverty, environmental destruction, corruption, and injustice.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

- If the United Nations could deploy heavily armed super-ethical peace-keeping robot armies into conflict zones to protect civilians and enforce ceasefires.
- If our towns and cities are policed by super-ethical robots that protect all citizens equally, 24x7, and never shoot our friends or family because of their race, religion, social class, lifestyle, or peaceful protesting.
- If super-ethical child-care robots accompany our children wherever they go, protecting them from danger, including physical, emotional, and sexual abuse.
- If ethically adequate corporations produce ethical products, provide ethical services, and pay ethical levels of unavoidable corporation tax.

Such a super-ethical society is possible; but only if we deliberately make it our goal, rise above polarizing politics, and act together in accordance with the undeniable truth that ethics are a higher power for good that transcends science, politics, nations, and religions.

5.6 Cyberanthropic Utopia

Because so many ludicrous utopias have been proposed that are just naïve science fiction fantasies, it is unsurprising that utopias have accumulated a bad reputation and are not taken very seriously. But it is shockingly common for apparently rational people to exhibit symptoms of classical conditioned-reflex (Pavlovian) negative responses to the stimulus word "utopia"; triggering emotional distress and bypassing their rational reasoning, as if any serious use of the word "utopia" has become a reputation-threatening scientific taboo.

However, now that artificial intelligence is making such impressive progress and showing no signs of slowing down nor having an upper-limit, Ross Ashby's 67-year-old prediction looks increasingly realistic: We might be able to "...solve our social problems by directing machines that can deliver an intelligence that is not our own."

Ashby's prediction hints at a possible definition of a minimum viable utopia:

A world where our social problems have been solved.

Utopia need not mean a "perfect" society, or that we all have flying cars, robot servants, and never have to go to work. Just fulfilling human needs and eliminating poverty would create a truly magnificent utopia. And as we start making progress achieving it, many other human problems, such as malnutrition, parasitic diseases, homelessness, hopelessness- and poverty-driven prostitution and crime will fade away and the world will become a very different and happier place to live in.

Let no one say it cannot be done. Ethically adequate societies have existed in the past, where resources were shared and the environment respected. What is new is that we can now do it synthetically, consciously, deliberately.

5.7 Third-Order Cybernetics

Since Heinz von Foerster made the distinction between first- and second-order cybernetics in 1974, many people have attempted to find a plausible definition for third-order cybernetics, but until now, no definition has gained acceptance.

This paper proposes that Third-Order Cybernetics should be defined as "the cybernetics of ethical systems", and that "the cybernetics of ethics", "Cybernetics 3.0", and "3oC" are all acceptable synonyms for it. Some of the supporting arguments for this proposal have already been mentioned, however a consolidated set of arguments are listed below:

1. Until now, second-order cybernetics (2oC) discussions about the need to create ethical systems, including the need for cybernetics itself to embody ethics, did not produce any satisfactory solution. Here, "satisfactory solution" is understood to mean something like the ethically adequate design process of Figure 3 that can be used systematically, for example by engineers, to create real systems that are ethically adequate. Recognizing this need but failing to fulfil the need could be referred to in the context of second-order cybernetics as "The Ethics Problem".

 doi:10.20944/preprints201907.0349.v1

2. The fact that Heinz von Foerster described "Ethics and Second-Order Cybernetics" as an "outrageous title" for a presentation, implies that ethics do not belong in the 2oC that he envisioned.
3. Whereas 2oC can be used for good or evil, the Ethical Regulator Theorem can only be used for good. This distinction is significant, and it also implies that ERT does not belong in 2oC. It is ERT's existence that creates the need to define 3oC.
4. If we extrapolate from first-order cybernetics having one observer and 2oC having a second observer, we would logically expect 3oC to introduce a third observer. This hypothesized third observer maps exactly onto the ERT requirement that ethical systems must have real-time integrity mechanisms that monitor and enforce compliance of the system with respect to an appropriate ethical schema.
5. An alternative, observer-free justification can be derived from the Good Regulator Theorem: A first-order cybernetic regulator requires a model of the system being regulated and a second-order cybernetic regulator can only achieve reflexivity by having a model of itself. Then to behave ethically, a third-order cybernetic regulator needs a third model, a model of acceptable (ethical) behavior, which is encoded in the ethics schema. It is then a consequence of the fact that every model requires observations as inputs, that brings into existence the need for observing part(s) to exist in the system. The need for these observations is independent of whether a cybernetician is watching or not. Whereas the observer-based argument of point 4 identifies no new requirements on the regulator, this model-based argument not only makes explicit that an ethical regulator requires three models, but it also requires observations as a direct consequence of using the models.
6. Together, ERT and the six-level framework (6LF) for classifying cybernetic and superintelligent systems (see Table 4) create a new paradigm that has greater explanatory and predictive power than 2oC. For example, producing the ERT effectiveness function, the Law of Inevitable Ethical Inadequacy, identifying that the ethical systems of cybernetic level 3 (CL3) are the missing type of cybernetic system that is necessary to integrate cybernetic systems and superintelligent systems into a common framework, explaining the impending bifurcation into either a cyberanthropic utopia or a cybermisanthropic dystopia (see Figure 6), and identifying deficiencies in capitalism. In addition, because 6LF integrates three classes of system that do not yet exist, it can help us navigate a rational path into the future, for example, by predicting the existence of a race condition and thus identifying a possible solution to the dangers that are posed by superintelligent machines. Such insights cannot be obtained using 2oC.
7. Ethical systems constitute a new type of system, and ERT + 6LF defines a new branch of cybernetics that goes beyond effectiveness, does not belong in 2oC, and solves "The Ethics Problem". Logically, the system that is created by joining the 2oC and ERT systems would be named third-order cybernetics.
8. Although Ranulph Glanville defined "the cybernetics of ethics and the ethics of cybernetics" as "cybernethics", this term is invented jargon that carries no meaning for people who are not familiar with its definition. By contrast, the term "third-order cybernetics" carries enough meaning for people who are familiar with the term "second-order cybernetics" to at least trigger interest and curiosity. Therefore, using the term "third-order cybernetics" instead of "cybernethics" has significant advantages, and enhances 6LF by increasing symmetry in Table 4.
9. Whereas it is impossible to define objectively which theories and practices belong in 2oC, thus making it an intimidating subject for outsiders to even contemplate mastering, ERT is defined and proved in eight pages and doesn't require any knowledge of 2oC. This means that the theorem, and how to apply it to systems of any type, can easily be taught to non-cyberneticians. It is therefore imperative that ERT + 6LF can make a fresh start as a new cybernetic speciality without being entangled with 45 years of fuzzy 2oC baggage. However, 3oC is not limited to ERT and 6LF, and will surely develop rapidly before it matures.
10. Unlike 2oC, 3oC has a fundamentally ethical purpose (making systems ethical) that together with the proposed grand challenge and the vision of a super-ethical society, create a unique

 doi:10.20944/preprints201907.0349.v1

opportunity for the cybernetics and systems sciences community to take a leading role in implementing a long-overdue and much needed systemic ethical revolution.

11. If someone makes the statement "I'm a second-order cybernetician", it reveals nothing about their ethics. But from now on, anyone who is brave enough to declare "I'm a third-order cybernetician" is making a bold assertion that they are part of the only scientific movement that is dedicated to making the world a better place.
12. Recognizing and declaring that 3oC is a new paradigm that is revitalizing and reunifying the 2oC community will help draw attention to the fact that something important and exciting is happening; a powerful attractor for energy and commitment. It is a genuine **ethical imperative** that we develop this new field for engineering, management, and the other sciences to use. And it is a task that cannot be performed by any of the more narrowly defined branches of science.

If these arguments are accepted, it is suggested that going forwards, the term second-order cybernetics should only be used to refer to the second-order cybernetics of effectiveness without the ethical, integrity, and transparency aspects, which belong to the 3oC layer that can be used to transform any effective but ethically inadequate system, such as a design process, second-order cybernetics, or capitalism, into an ethical system.

Many people thought that cybernetics had faded away after peaking in the 1960s or early 1970s, but that peak was just a local maximum: Cybernetics is rebooting as Cybernetics 3.0, and this time it's going to be harder to ignore, because we'll be applying requisite Purpose, Truth, Variety, Predictability, Intelligence, Influence, Ethics, Integrity, Transparency — and Love; because a sincere desire to make the world a better place emerges ***only*** in people who love humanity and the biosphere unconditionally.

By contrast, deep down, non-empaths only care about themselves or members of their own nation, race, religion, family, or gang, and consequentially embody a conflict of interests that compels them to act against the greater good.

5.8 Our Future Epilog or Eulogy

We are approaching a decisive fork in the road in the evolution of 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 humanity politically and economically.

We are the *only* generation that has the chance to steer the fate of future generations of humanity towards being collectively ruled, potentially for eternity, by benevolent super-ethical systems that create a stable cyberanthropic utopia for us, effectively and ethically minimizing human suffering and environmental problems, rather than allowing hubris and super-unethical systems to either enslave most of us in a cybermisanthropic dystopia or cause the extinction of our species to become a footnote in Gaia's geological record.

5.9 The Path Forwards

To start steering the future of humanity and our wonderful planet towards becoming a stable cyberanthropic super-ethical society, this paper proposes establishing an independent, non-profit institute with ambitious goals that lie in the areas of research, development, standards, certification, legislation, and democracy.

5.9.1 Research and Development

The institute will promote theoretical and practical progress:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2019 doi:10.20944/preprints201907.0349.v1

5.9.2 Standards and Certification

The 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.

5.9.3 Legislation and Democracy

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

- 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 explicitly excluded from requiring certification.
- Require that all new systems and processes are designed to be ethically adequate.
- Extend political representation to every member of society by giving parents proxy votes to cast on-behalf of their children who are too young to vote, but not too young for morally bankrupt politicians to load up with unsustainable debt liabilities, while underfunding the public education system and allowing unethical corporations to maximize short-term profits by devastating the environment for all future generations of humanity. What we currently call "universal suffrage" [21] is a perversion of the true meaning of the word "universal".

5.10 Example: Applying ERT to 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 identify ways to improve each ethical requisite as it applies to yourself. Table 5 illustrates how you can use ERT to make yourself a better ethical regulator. Or, stated in ERT algebra:
MakeEthical(yourself) = Table 5

Table 5. Ways to become a better ethical regulator

Requisite	Example set of self-improvement interventions
Purpose	To clarify your purpose in life and help you to recognize your strongest motivating thoughts, write down your most important life goals: 1. 2. 3. 4. 5.
Truth	To become a good judge (effective and ethical) of who tells the truth and who distorts it, seek alternative information sources that are genuinely independent of your primary sources. Investigate any inconsistencies that you notice, modify the reputation of liars, and resolve to always doubt them skeptically in future.
Variety	Brainstorm new actions, responses, and strategies that you have never previously considered, to make progress towards achieving your goals.

 doi:10.20944/preprints201907.0349.v1

Predictability	Improve your model of human behavior by studying the following Wikipedia articles until you are competent at recognizing the patterns in yourself and others: • List of cognitive biases [22] • Defence mechanisms [23] • List of fallacies [24] • Demagogue [25]
Intelligence	Take a course or read a book on critical thinking or personal effectiveness.
Influence	Identify ways that you can increase your influence (on your family, friends, colleagues, clients, or society) to achieve your life goals and promote your ethical values.
Ethics	Write down five undesirable, unethical, or disrespectful behaviors that, up until now, you have tolerated in other people, organizations, or corporations: 1. 2. 3. 4. 5. Next to them, write down five undesirable, unethical, or disrespectful behaviors that, up until now, you have tolerated in yourself. If you can't think of five things about yourself, read the Wikipedia article: Denial [26]. 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.
Integrity	Seek to stop or prevent all the undesirable, unethical, or disrespectful behaviors 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 for Purpose and Ethics until they genuinely reflect who you are and how you want your world to become.

5.11 Ethically Resonant Wisdom

If you distil different solutions that contain alcohol, you get pure alcohol that is free of impurities. And if you distil different religions and philosophies that contain ethical wisdom, you get pure ethical wisdom that is free of culturally-specific dogma. Such ethical wisdom is universal, and resonates with all good people, regardless of their worldview, politics, nationality, or religion.

And because pure 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. For example, consider the following selected quotes:

Mahatma Gandhi (1869-1948):

1. The future depends on what you do today.
2. Be the change you wish to see in the world.
3. The difference between what we do and what we are capable of doing would suffice to solve most of the world's problems.
4. 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.
5. First they ignore you, then they laugh at you, then they fight you, then you win.

 doi:10.20944/preprints201907.0349.v1

6. Happiness is when what you think, what you say, and what you do are in harmony.
7. Non-cooperation with evil is as much a duty as is cooperation with good.
8. Poverty is the worst form of violence.
9. Capital as such is not evil; it is its wrong use that is evil.
10. There is sufficiency in the world for man's need,
    but not for man's greed.
11. There are people in the world so hungry,
    that God cannot appear to them except in the form of bread.
12. Those who say religion has nothing to do with politics do not know what religion is.
13. Where love is, there God is also.
14. God has no religion.
15. There is a higher court than the courts of justice and that is the court of conscience.
16. They may torture my body, break my bones, even kill me.
    Then they will have my dead body, but not my obedience.
17. Victory attained by violence is tantamount to a defeat, for it is momentary.
18. 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?
19. 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.

His Holiness Pope Francis:

20. 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.
21. Hatred is not to be carried in the name of God. War is not to be waged in the name of God!
22. Human rights are not only violated by terrorism, repression, or assassination,
    but also by unfair economic structures that create huge inequalities.
23. 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.
24. 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.
25. Women in the church are more important than bishops and priests.
26. All that is good, all that is true, all that is beautiful, God is the truth.
27. We all have the duty to do good.
28. 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.

His Holiness the Dalai Lama XIV:

29. All religious institutions, despite different philosophical views,
    all have the same message — a message of love.
30. If you can, help others; if you cannot do that, at least do not harm them.
31. The whole purpose of religion is to facilitate love
    and compassion, patience, tolerance, humility, and forgiveness.
32. 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.
33. The ultimate authority must always rest with the individual's own reason and critical analysis.
34. The true hero is one who conquers his own anger and hatred.
35. 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.
36. A lack of transparency results in distrust and a deep sense of insecurity.

 doi:10.20944/preprints201907.0349.v1

37. In our struggle for freedom, truth is the only weapon we possess.
38. Where ignorance is our master, there is no possibility of real peace.
39. Through violence, you may "solve" one problem, but you sow the seeds for another.
40. 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.
41. A truly compassionate attitude toward others does not change even if they behave negatively or hurt you.
42. I defeat my enemies by making them my friends.
43. When you practice gratefulness, there is a sense of respect toward others.
44. With realization of one's own potential and self-confidence in one's abilities, one can build a better world.
45. If you think you are too small to make a difference, try sleeping with a mosquito.
46. 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.
47. The ultimate source of happiness is not money and power, but warm-heartedness.
48. The more you are motivated by love, the more fearless and free your action will be.
49. Love and compassion are necessities, not luxuries.
    Without them humanity cannot survive.
50. Love is the absence of judgement.
51. Be kind when possible. It is always possible.

Dr. Martin Luther King Jr. (1929-1968):

52. We must discover the power of love, the power, the redemptive power of love.
    And when we discover that we will be able to make of this old world a new world.
    We will be able to make men better.
    Love is the only way.
53. I say to you, "I love you. I would rather die than hate you."
    And I'm foolish enough to believe that through the power of this love, somewhere, men of the most recalcitrant bent will be transformed.
54. Darkness cannot drive out darkness; only light can do that.
    Hate cannot drive out hate, only love can do that.
55. Those who love peace must learn to organize as effectively as those who love war.
56. True peace is not merely the absence of tension.
    It is the presence of justice.
57. Injustice anywhere is a threat to justice everywhere.
58. In a real sense, all life is inter-related.
    All men are caught in an inescapable network of mutuality, tied in a single garment of destiny.
    Whatever affects one directly, affects all indirectly.
59. Every man must decide whether to walk in the light of creative altruism or in the darkness of destructive selfishness.
60. The time is always right to do the right thing.
61. 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.
62. You are not only responsible for what you say, but also for what you do not say.
63. Our lives begin to end the day we become silent about things that matter.
64. Our scientific power has outrun our spiritual power.
    We have guided missiles and misguided men.

 doi:10.20944/preprints201907.0349.v1

65. A nation that continues year after year to spend more money
on military defense than on programs of social uplift is approaching spiritual doom.
66. We should never forget that everything Adolf Hitler did in Germany was “legal”
and everything the Hungarian freedom fighters did in Hungary was “illegal”.
67. 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.
68. 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.

Nelson Mandela (1918-2013):

69. 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.
70. 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.
71. Overcoming poverty is not a gesture of charity.
It is an act of justice.
72. As long as poverty, injustice and gross inequality persist in our world,
none of us can truly rest.
73. Education is the most powerful weapon which you can use to change the world.
74. It is in your hands to create a better world for all who live in it.
75. May your choices reflect your hopes, not your fears.

Albert Einstein (1879-1955):

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

Margaret Mead (1901-1978):

77. Never doubt that a small group of thoughtful, committed citizens can change the world;
indeed, it's the only thing that ever has.

Bertolt Brecht (1898-1956):

78. Change the world, she needs it.

Percy Bysshe Shelly (1792-1822):

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

Leonardo da Vinci (1452-1519):

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

Despite the authors of these quotes being separated by space, time, and their affiliations, it's 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 all came together in one room to plan an ethical revolution to make the world a better place.

5.12 The Law of Unethical Arguments

It is a certainty that all good people (without exception) are supportive of redesigning unethical systems, organizations, corporations, products, taxes, laws, regulations, and processes to make them more ethical. And it makes sense that the only people who want such systems to remain unethical and vulnerable to tampering and abuse are the small minority of people who benefit (directly or indirectly) from those systems remaining unethical.

The final law in this manifesto for a nonviolent global ethical revolution to create a stable cyberanthropic super-ethical society is defined in one sentence:

"Because no ethical argument can exist against making a system ethical,
anyone who argues against this goal,
obstructs progress towards this goal,
or abuses its sincere supporters,
is either objectively unethical, corrupt, or evil."

□

6 Conclusions

The Ethical Regulator Theorem creates a theoretical basis for applied ethics, enabling designers to systematically evaluate, improve, and design ethically adequate systems. Because it is a universal theorem that can be applied to any system, the possible areas of application are vast and potentially world-changing.

The six-level framework for classifying cybernetic and superintelligent systems leads to a theory-based solution to the danger that superintelligent machines might cause a dystopia: We *must* create ethical systems *before* we create superintelligent systems!

By creating a well-defined decision function (IsEthical) that identifies systems as being either ethically adequate or ethically inadequate, ERT provides a semantic precision that avoids the ambiguities and unstated assumptions that multiply exponentially when the word "ethical" is bandied around as if we all understand it to mean the same thing. But "ethical AI", "ethical product", and "ethical corporation" can mean very different things to different people. By contrast, ERT gives terms like "ethically adequate AI", "ethically adequate product", and "ethically adequate corporation" a much more precise meaning, and could even be made the subject of a formal certification process that qualifies recipients to use an Ethically Adequate branded logo and reduce their liability insurance premiums.

ERT's universality means that the nine dimensions define an abstraction layer that can be mapped onto the regulators of any systems in any domain, thus enabling communication and learning to take place between experts in seemingly unrelated domains.

Because of the flaw that was identified in Heinz von Foerster's Ethical Imperative, a new definition is proposed, which is intended to embody both the essence of the proposed grand challenge and a principle for good that is universal and worthy of the magniloquent name "Ethical Imperative":

"**Always strive to make new and existing systems ethically adequate!**"

The proposed grand challenge to implement a systemic ethical revolution is neither a new religion nor a political movement, it is a response to Johann Eder's call for a grand challenge in Vienna [27] and Irma Wilson and Pamela Buckle Henning's call to action for the systems sciences community in Berlin [28].

This ethical revolution is the product of a compassionate heart and mind, employing the Ethical Regulator Theorem to generate maximally coherent ethical interventions in multiple complex systems, such as the computational, corporate, criminal, cybernetic, personal, political, product development, psychological, scientific, social, and spiritual [29] realms. And all such interventions resonate, not only with each other, but also with all good people who have ever existed — or ever will.

This revolution is long overdue, and we are privileged to live in these exciting times, but passively watching from the sidelines, or doing nothing, only helps the criminals, psychopaths, demagogues, and ethically indifferent corporations to create and exploit the pathological chaos and emergent problems that, until now, we have accepted as normal. It's time for all good people to make a commitment to yourself to do everything that you can to research, design, educate, campaign, love, heal, and fight for a better world.

"To be bold enough to consciously and deliberately reach beyond ourselves,
to accept a grand challenge for the greater good,
would be an act of self-actualization."
— Stella Octangula [30]

Just like we have legislation and non-negotiable expectations that passenger aircraft are designed to include expensive redundant subsystems to avoid having single points of failure in flight-safety-critical systems, and that all electrical products that we purchase conform to strict safety standards, we must change our attitudes, to create a cultural shift that makes it totally unacceptable and utterly unthinkable to knowingly design systems or sell products that are ethically inadequate. Outrage at such behavior is appropriate.

We must demand strict legislation and higher standards to force ethically indifferent corporations to stop their races to the bottom and cost externalization strategies. In truth, only certifiably ethical corporations can be trusted to produce ethically adequate products and services that help to make the world a better place for the entire human race.

Arguably, the root cause of all evil is a lack of ethics, and by systematically applying the Ethical Regulator Theorem, we can reliably increase ethical behavior in many classes of systems; progressively reducing unethical behavior, reducing unethical suffering, and setting a course for humanity towards the tipping-point where we will experience a peaceful social phase-transition to a stable cyberanthropic super-ethical society.

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 super-ethical systems that can create a better world, for the need is great.

Conflicts of Interest: The author declares no conflicts of interest

References

1. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.
2. Conant, R.C.; Ashby, W.R. Every good regulator of a system must be a model of that system, INT J SYST SCI **1970** 1(2), 89-97.
3. Ashby, W.R. (1956). An Introduction to Cybernetics, Chapman and Hall, London, 1956. Available online: rossashby.info/Ashby-Introduction-to-Cybernetics.pdf (accessed on 18 July 2019).
4. Cutler, A.; Pribić, M.; Humphrey, L. Everyday Ethics for Artificial Intelligence: A practical guide for designers and developers, IBM Corporation, 2018. Available online: www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf (accessed on 18 July 2019).
5. European Commission, High-Level Expert Group on Artificial Intelligence: Draft Ethics Guidelines for Trustworthy AI, 2018. Available online: ec.europa.eu/digital-singlemarket/en/news/draft-ethics-guidelines-trustworthy-ai (accessed on 18 July 2019).
6. von Foerster, H. On Constructing a Reality. In *Environmental Design and Research*, Volume 2, pp 35-46. F.E. Preiser (ed.), Dowden Hutchinson and Ross, Stroudsburg PA, 1973.
7. von Foerster, H. Ethics and Second-Order Cybernetics. In: *Understanding Understanding*. Springer, New York, NY, 2003.
8. Beer, S. *Brain of the Firm*, John Wiley and Sons, New York, 1972.
9. Asimov, I. Runaround, in Astounding Science Fiction, March 1942.

10. Ashby, M. How to apply the Ethical Regulator Theorem to crises, Acta Europeana Systemica, n°08, Brussels, Belgium, 2018. Available online: aes.ues-eus.eu/aes2018/aes806_Mick-Ashby.pdf (accessed on 18 July 2019).
11. Stevens, J.P. Opinion of Stevens J., Supreme Court of the United States. Citizens United, Appellant v. Federal Election Commission, *Legal Information Institute*, Cornell University Law School, 2010. Available online: www.law.cornell.edu/supct/html/08205.ZX.html (accessed on 18 July 2019).
12. Thomas, D.R.; Beresford, A.R.; Rice, A. Security Metrics for the Android Ecosystem, 2015. Available online: www.cl.cam.ac.uk/~drt24/papers/spsm-scoring.pdf (accessed on 18 July 2019).
13. Umpleby, S.A. What comes after second order cybernetics, *Cybernetics and Human Knowing* **2001**, 8(3):87-89.
14. Kuhn, T. *The Structure of Scientific Revolutions*, University of Chicago Press, 1962.
15. Glanville, R. The Cybernetics of Ethics and the Ethics of Cybernetics, Tutorial at 21st American Society for Cybernetics Conference, Virginia Beach, USA, 1986.
16. Turing, A.M. Computing Machinery and Intelligence, *Mind* **1950**, 59:433-460.
17. Ashby, W.R. Advanced society planned as a super brain, *Journal*, volume 14, 1951, page 3528. Available online: rossashby.info/journal/page/3527.html (accessed on 18 July 2019).
18. Ashby, W.R. Power and I.Q. have many similar properties, *Journal*, volume 16, 1952, pages 4276-4278. Available online: rossashby.info/journal/page/4276.html (accessed on 18 July 2019).
19. Ashby, W.R. Appearance of a super-clever machine, *Journal*, volume 16, 1952, pages 4279-4280. Available online: rossashby.info/journal/page/4279.html
20. Clarke, A.C. Hazards of Prophecy: The Failure of Imagination, in *Profiles of the Future: An Inquiry into the Limits of the Future*, 1962.
21. Wikipedia: Universal suffrage. Available online: wikipedia.org/wiki/Universal_suffrage (accessed on 18 July 2019).
22. Wikipedia: List of cognitive biases. Available online: wikipedia.org/wiki/List_of_cognitive_biases (accessed on 18 July 2019).
23. Wikipedia: Defence mechanisms. Available online: wikipedia.org/wiki/Defence_mechanisms (accessed on 18 July 2019).
24. Wikipedia: List of fallacies. Available online: wikipedia.org/wiki/List_of_fallacies (accessed on 18 July 2019).
25. Wikipedia: Demagogue. Available online: wikipedia.org/wiki/Demagogue (accessed on 18 July 2019).
26. Wikipedia: Denial. Available online: wikipedia.org/wiki/Denial (accessed on 18 July 2019).
27. Eder, J. Grand Challenges for Computer Science Research: Ross Ashby Memorial Lecture of the International Federation for Systems Research, in *Cybernetics and Systems* 2010; Trappl, R. (Ed.), Vienna, Austria, pp. xix-xxv.
28. Wilson, I.; Buckle Henning, P. A Call to Action for the Systems Sciences Community, in Proceedings of the 59th Annual Meeting of the International Society for the Systems Sciences, Berlin, Germany, 2015. Vol. 1(1). Available online: journals.isss.org/index.php/proceedings59th/article/viewFile/2609/836 (accessed on 18 July 2019).
29. Wilson, T.A. The Ethical Regulator Theorem, *YouTube*, 2017. Available online: youtube.com/watch?v=NLhUajpMOI4 (accessed on 18 July 2019).
30. Octangula, S. Structure, Environment, Purpose, and a Grand Challenge for the ASC, 2011. Available online: asc-cybernetics.org/CofC/wp-content/uploads/2011/02/StructureEnvironment-Purpose-and-a-Grand-Challenge-for-the-ASC-V2.0.pdf (accessed on 18 July 2019).