PROBLEMS WITH ABSTRACT OBSERVERS AND ADVANTAGES OF A MODEL-CENTRIC PARADIGM

Mick Ashby

Fellow of the Cybernetics Society: http://cybsoc.org
Archivist of the W. Ross Ashby Digital Archive: http://rossashby.info
Former Trustee of the American Society for Cybernetics: https://asc-cybernetics.org
Email: ethics@ashby.de
Twitter: @AshbyEthics

Keywords: ethical AI, ethical robots, philosophy of cybernetics, third-order cybernetics, empathy, sapientocracy

ABSTRACT

Since 1974, when Heinz von Foerster made the distinction between "the cybernetics of observed systems" as first-order cybernetics (1oC) and "the cybernetics of observing systems" as second-order cybernetics (2oC), cybernetics has been dominated by this observer-centric paradigm. However, the abstract conceptualization of a pure observer has no purpose except observing, which is like a person who is meditating so deeply with open unfocussed gazing eyes that they neither react to what they observe nor make memories of it. But when the observer has a specific purpose, "observer" ceases to be the best description of them. For example, if their purpose is to create a model of the system that is observed, then the best name for them is "modeler", or if they are observing the system to ensure that it breaks no laws, then the names "auditor" or "conscience" are more accurate and useful descriptions than "observer". And because there are an infinite number of possible reasons why an observer might observe a system, the term "observer" is completely ambiguous with respect to purpose, which renders it deficient in clarity and utility. So, when von Foerster introduced "the cybernetics of observing systems" it was vulnerable to many philosophical interpretations that are of no practical use. Consequentially, 1974 can be regarded as the year of the schism of cybernetics into two very different communities, which can be characterized (without observers) as "the cybernetics of regulated systems" and "the philosophy of cybernetics". This schism has been as significant as the distinction between 1oC and 2oC. Just as the boundary of the system being considered can be decided arbitrarily as seems most useful, the distinction of whether the observed system contains an observer (2oC) or not (1oC) is quite arbitrary because there are many other equally valid criteria that can be used to slice the field of cybernetics into two parts. Yet when a cybernetician is deciding what the boundaries are of what she will consider to be the system being studied or its environment, she strives to maximize clarity and utility. However, this is not true of von Foerster's observer-centric 1oC/2oC paradigm, and by his own account, it also limits cybernetics by excluding the possibility of the existence of a meaningful definition of third-order cybernetics. Fortunately, we can define a model-centric paradigm that has more utility to the practical application of cybernetics to real systems. In the model-centric paradigm, a first-order cybernetic regulator needs a model of the system that is being regulated (as required by the good regulator theorem), and because every model requires observations as inputs, the model not only brings into existence the need for an observer to exist, it also defines very precisely what information that observer must collect. A second-order cybernetic regulator maintains a second model, a model of itself, which is the only way it can achieve reflexivity. The second model requires self-observations, which requires a second observer that is well-

defined by the needs and purpose of the second model. Now we are ready to imagine a type of regulator that requires a third model and a third observer: If the third model is a model of acceptable (ethical[1]) behaviour, then a third observer is a necessary element of the system's "conscience" that prevents or detects any violations of the model of ethical behaviour. In this paradigm, the cybernetics of systems that are designed to exhibit ethical behaviour can be characterized as third-order cybernetics (3oC). By being able to extend the paradigm to include ethical systems, the model-centric paradigm brings clarity and utility that is not possible using the philosopher-friendly observer-centric paradigm and its under-specified (abstract) observers.

INTRODUCTION

In 1948, Norbert Wiener defined cybernetics as "the scientific study of control and communication in the animal and the machine" [1]. This defines a scope for cybernetics that includes all types of regulators and communicators. And if we accept that philosophy is not science, then Wiener's original definition of cybernetics excludes any philosophical aspects.

In 1974, Heinz von Foerster [2] made the distinction between "the cybernetics of observed systems" as first-order cybernetics (1oC) and "the cybernetics of observing systems" as second-order cybernetics (2oC). Since 1974, the cybernetics community has been dominated by this observer-centric paradigm.

Although he might have originally intended it purely as a distinction between two approaches to performing cybernetics, it has been widely generalized as a model for understanding different types of systems. Thus, endowing meaning to terms like first-order system and second-order system, as illustrated by Ranulph Glanville "The question arises about differences in how we observe a system that is first or second order." [3].

However, von Foerster's definitions use the terms "observed" and "observing" without being more specific. Of course, there will be an underlying purpose behind any observations, but the abstract conceptualization of a pure observer has no purpose except observing, which is like a person who is meditating so deeply with open unfocussed gazing eyes that they neither react to what they observe nor make memories of it. And when the observer has a specific purpose, "observer" ceases to be the best description of them.

For example, if the purpose of an observer is to create a model of the system that is observed, then the best name for them is "modeler", or if they are observing the system to ensure that it breaks no laws, then the names "auditor" or "conscience" are more accurate and useful descriptions than "observer". And because there are an infinite number of possible reasons why an observer might observe a system, the term "observing" is completely ambiguous with respect to purpose, which renders it deficient in clarity and utility.

Because the expression "observing systems" is devoid of actual purpose, von Foerster's definition of 2oC could be rewritten as "the cybernetics of X, where X is any activity that requires observations to be made". Stated in this way, the extreme ambiguity of his definitions

[1] Throughout this paper, an ethical system is equated with one that is internally constrained to only exhibit behaviour that is defined as being acceptable in the culture within which the system exists. If we accept that a society's laws, regulations, and rules are its operative best approximation to that society's most important ethical principles, then what this paper refers to as an ethical regulator can be equated with a strictly law-abiding regulator. Thus, the term ethical behaviour is not intended to include interpretations such as ethical reasoning, ethical questioning, and discourse.

becomes more obvious, and Mead's "cybernetics of cybernetics" is only one of many possible interpretations.

Therefore, when von Foerster introduced 2oC as "the cybernetics of observing systems" it had a multitude of possible interpretations that went beyond Wiener's precisely defined "scientific study of control and communication" to include just about anything, such as "thinking about thinking" and "understanding understanding", which have been a consideration of philosophers dating back to the ancient Greeks and continues today in philosophy departments around the world. This encouraged a profound intellectual inclusiveness that is not enjoyed by any other science, but it is not without drawbacks.

Consequentially, 1974 can be regarded as the year of the schism of what was originally Wiener's science of cybernetics into two very different communities, which can be characterized (without observers) as "the cybernetics of regulated systems" and "the philosophy of cybernetics". This schism has been as significant as the distinction between 1oC and 2oC. Whereas Kline [4] and Scott [5] considered it and other disunities from a historical perspective, this paper is more concerned with the implications and limitations of von Foerster's observer-centric paradigm.

Although this paper takes a position that must be contrasted with von Foerster's definitions and interpretations, which are characterized as being philosophical, it is not intended to suggest that philosophical aspects are less worthy, totally unscientific, or that they should be excluded from cybernetics, but rather that cybernetics embodies both scientific and philosophical approaches, and that the more philosophical interpretation of von Foerster's 1oC/2oC definitions as two different ways of performing cybernetics has arguably dominated the equally valid possible scientific interpretation of 1oC and 2oC as increasing orders of regulation. We explore the implications of the second interpretation, and find that it produces valuable results and has practical implications for AI and robotics that cannot be derived from the observer-centric interpretation. The use of these characterizations of one as being philosophical and the other as being scientific is not intended to deny that there is a significant complementary overlap between both approaches.

DIVERSITY AND DUALISM

Just as the boundary of the system being considered by a cybernetician can be decided arbitrarily as seems most useful, the distinction of whether a system contains an observer (2oC) or not (1oC) is quite arbitrary because there are many other equally valid criteria that can be used to slice the field of cybernetics into two parts. Yet when a cybernetician is deciding what the boundaries are of what she will consider to be the system being studied or its environment, she strives to maximize clarity and utility.

However, this is not true of von Foerster's observer-centric 1oC/2oC paradigm. By arbitrarily introducing observers into his definitions that were not present in Wiener's definition, clarity and utility have not been maximized, and suddenly everyone felt entitled to offer their own alternative definitions of 1oC and 2oC, as described by Stuart Umpleby [6]. Table 1 is based on a table from Umpleby, but with entries added for Wiener and Mead to provide more context for this discussion.

Table 1: Definitions of cybernetics

Author	First-Order Cybernetics	Second-Order Cybernetics
Wiener	The scientific study of control and communication	
Mead	Cybernetics	The cybernetics of 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

This diversity of definitions reflects that while some cyberneticians remained focussed on the practical scientific study of real systems, others entered the realms of philosophy, for example, exploring the circularity of Margaret Mead's "cybernetics of cybernetics" and radical constructivism, just as 20 years earlier, on page 4814 of his journal, Ross Ashby had reflected on the circularity of his brain studying "the brain" and identified a significant limitation of it:

"For some time, I have had the hunch that there is a profound circularity in science. Since I have to use a mammalian brain to think with, I am likely to find in the world only what it lets me find. Ultimately, as the 'world as I see it' is analysed, I shall find simply that I am looking, as it were, at the underside of my own cortex." [7]

From the point-of-view of cybernetics as a science, interpreting second-order cybernetics as "the cybernetics of cybernetics" rather than "the cybernetics of reflexive systems", of which "the cybernetics of cybernetics" is merely a special case, can be viewed as a decision that established a philosophy of cybernetics that has dominated the cybernetic narrative and has somewhat attenuated the scientific study of the cybernetics of reflexive systems.

This paper's criticism of von Foerster's 1oC/2oC paradigm is not intended to suggest that it has not been useful. Even if it has weaknesses, it made an important distinction that helped cybernetics develop into what it is now. But after more than 45 years as the dominant paradigm in cybernetics, we suggest that it has probably contributed pretty much all that it ever can, and that it is time to find a new paradigm that can bring cybernetics firmly back into the realms of science.

The practitioners of the philosophy of cybernetics have certainly added an incredible richness to the field of cybernetics, but it has been at the cost of making cybernetics appear to be less scientific. And it is clear that topics such as "thinking about thinking", "understanding understanding", and radical constructivism represent significant shifts towards philosophy and away from cybernetics as a science. But these days, the world is in such a terrible mess that what we desperately need is better cybernetics, not better philosophy.

The Possibility of Third-Order Cybernetics

In a 1990 interview, von Foerster categorically asserted that there is no possibility of the existence of a meaningful definition of third-order cybernetics:

"… it would not create anything new, because by ascending into 'second-order,' as Aristotle would say, one has stepped into the circle that closes upon itself. One has stepped into the domain of concepts that apply to themselves." [8].

Ranulph Glanville went further to claim that a third-order system cannot exist because it collapses into being equivalent to a first-order system [3].

Despite these claims by von Foerster and Glanville that defining third-order cybernetics would not create anything new, many people have attempted to do exactly that, by extrapolating from von Foerster's definitions of 1oC and 2oC to define 3oC. The result has been various competing intelligent and creative proposals for possible definitions of 3oC that are merely plausible. None have had arguments that were sufficiently compelling to gain widespread acceptance.

The Model-Centric Cybernetics Paradigm

Rather than trying to extrapolate from von Foerster's observer-centric definitions, we shall construct an alternative paradigm from first principles, seeking to maximize clarity, precision, and utility, completely avoiding any reliance on the problematic concept of abstract observers, and above all making the new paradigm rigorously scientific such that, if he were still alive today, Wiener might have given it his approval.

If we start with Wiener's original definition of cybernetics as "the scientific study of control and communication", from the word "control" we can infer that it includes all types of regulators, R, and we can assume that they have a purpose, P, to regulate an entity that we can refer to as the regulated system, S.

For completeness, we must also consider Wiener's use of the word "communication". If a communication has any effect on the recipient, we can regard it as a special type of control. And if it has no effect on the recipient or if the communication reaches no receiver, then it is still equivalent to an act of control but it corresponds to the special case of a non-action.

First-Order Regulators: Simple Regulators

Let us define a first-order cybernetic regulator, R_1, that has a purpose, P_1. In order for it to be effective, R_1 must have a model, M_1, of the regulated system, S. This is a direct consequence of the Conant-Ashby good regulator theorem, which proved that "every good regulator of a system must be a model of that system" [9]. We know that M_1 is a model of S, but the scope and precision of M_1 are very much determined by P_1.

The effect of purpose on the scope and detail of the model can be illustrated by the fact that in a conceptual model of a pig brain, the complexity and level of detail must be considerably higher for a researching neurologist to be effective than is required by a butcher. They both require models of the same system, but it is the difference in their purposes that requires very different actual models.

Next, because a fundamental characteristic of a model is that it requires observations as inputs, the model not only brings into existence the need for an observer, O_1, to exist in R_1, but that observer is not an unconstrained abstract observer that might observe anything (or everything) because the inputs of M_1 define very precisely what information O_1 must collect.

However, having a purpose, P_1, appropriate model, M_1, and our well-defined observer, O_1, is insufficient to realize the functionality of an effective regulator. We must introduce an entity that is capable of making appropriate decisions to select the best actions (or communications) to achieve effective regulation. This decision-making unit must embody some kind of intelligence that could be biological, algorithmic, or baked into the design by the intelligence of the designer. Because we require clean definitions, this intelligence cannot be attributed to the observer being an intelligent observer, it must be identified as a separate component of R_1, which we shall name I_1. Finally, the regulator requires a channel, C_1, that transmits the selected action or communication to the regulated system.

Now we have a complete definition that a first-order cybernetic regulator, R_1, consists of a purpose P_1, model, M_1, observer, O_1, intelligence, I_1, and control channel, C_1. And Heinz would surely not hesitate to confirm that such a regulator conforms to his understanding of a first-order regulator that can be designed using 1oC.

It is worth highlighting that we refer to von Foerster's 1oC/2oC paradigm as being observer-centric because the observer takes a primary role in his definitions, as if everything revolves around the cybernetician and his observations, which has possibly biased cyberneticians to only think of observers as being intelligent observers, like themselves. But in the model-centric paradigm, the observer is arguably the least important element in the regulator because it serves simply as a data source to obtain information to feed the model. And the model is determined by the purpose of the regulator. So, it is fair to say that O_1 is to M_1 and I_1 as a keyboard is to a program and a computer. Observers and keyboards are necessary, but they are neither the most interesting nor the most important parts.

Second-Order Regulators: Reflexive Regulators

Our next step is to consider reflexive regulators that have to take themselves into account, such as the U.S. Federal Reserve's Federal Open Market Committee (FOMC), which generally meets every six weeks to decide whether to change the U.S. central bank's base interest rate. In addition to needing economic models in M_1, it can only be effective if it also has self-knowledge about how long it is until the next meeting, which is its next scheduled opportunity to make another interest rate change. If the FOMC decided that the next meeting would be in six months instead of the normal six-week cadence, this self-knowledge must be taken into account and affect their decision making and some parameters of the economic simulations that they run.

Another example is that a two-armed robot only requires a 1oC regulator to control the arms to perform simple tasks. But implementing reflexivity could mean that if it detects that one arm is non-functional, it would try to achieve its goals using just the one arm that works.

So, a second-order cybernetic regulator, R_2, not only includes all of R_1, but it must also have a self-referential purpose, P_2, of taking itself into account, and maintain a second model, M_2, which captures key aspects about itself. This model requires self-observations that must be performed by a second observer, O_2, that is well-defined by the needs of the second model. And it must also include some degree of intelligence, I_2, and a control channel, C_2.

Another example is that because of our constant subconscious self-monitoring of our health and vitality, if we find ourselves in a situation where we need to jump over an object, our M_2 model can, at any moment, provide a sense of self-knowledge of how high and how far we are confident of being able to jump. When we are ill or tired, knowing that we cannot jump a stream or a log could be important for our survival. The model of Newtonian physics that allows us to

instinctively know how to jump effectively has been learned through trial and error, and is stored in our M_1 model of how to interact effectively with our environment. So I_2 uses M_2 to provide potentially complex reflexive information to I_1 via channel C_2, which implies that in humans and animals, the nature of C_2 is electrical.

If the M_2 model can provide real-time information about what R is capable of in terms of possible actions and communications, then it is actually the source of R's knowledge of its own **variety**. Much has been made of the law of requisite variety, and people often talk of having requisite variety as if it were some sort of magical deus ex machina, but the M_2 model appears to be where information about possible variety of the regulator is stored.

Again, Heinz and second-order cyberneticians would surely agree that the regulator $R_2(R_1(S))$ conforms to their concept of a second-order (reflexive) cybernetic regulator that can be designed using 2oC.

In humans, there are fuzzy boundaries between models, like when someone is learning to drive a car with manual gear shifting. Initially they are very conscious of their hands, feet, pedals, gear-stick, and looking into mirrors. But with time and practice the ability to drive gradually migrates into an instinctive subconscious ability. Thus, the optimizations and black-box nature of human minds create limits for being able to use human examples to confirm or illustrate the theory of discrete orders of regulation and models. These limitations do not apply to robots.

Of course, it can be argued that the M_1 model could be extended to include all the reflexive information that we propose is encoded in the M_2 model, but that is an implementation optimization that ignores that reflexivity is a significant and important difference that deserves to be treated separately. If an existing regulator is repurposed to regulate a different type of system, it would be disadvantageous if the original implementation did not clearly distinguish between the environment and the regulator. Self-knowledge is a different category of information. For example, if we recognize that we do not have sufficient variety to control a system adequately, we can perform a strategic reflexive self-evaluation to identify which of our abilities we can improve to increase our possible variety to improve the quality of our performance as a regulator, and such improvements are limited to variables that belong to the M_2 model. And when we consider regulators such as AI and robots, the regulator must have a clear distinction between what is "self" that is subject to essentially deterministic direct control or knowledge of one's own limitations of possible variety and things in its environment that are interacted with and require probabilistic predictions of their future states, values, positions, or behaviour. So M_2 provides self-knowledge about variety and M_1 is used to make predictions.

Third-Order Regulators: Ethical Regulators

Because we are exploring a model-centric paradigm, rather than an observer-centric paradigm, we are now ready to imagine a type of regulator, R_3 that includes all of R_2, but also requires a third type of model.

This is a key difference from previous approaches that have tried to extend von Foerster's observer-centric paradigm by adding a third observer on top of the existing two observers, which can lead people into philosophical interpretations that ignore real systems.

But there is no problem adding a third model in the model-centric paradigm because although R_3 can be viewed as a speculative extrapolation, it is building on top of our purposeful R_2, which is defined more precisely and rigorously than von Foerster's 2oC.

If a third-order regulator, R_3 has a purpose P_3 of constraining $R_2(R_1(S))$ to exhibit only ethically acceptable behaviour, then R_3 must have a model, M_3, of what constitutes acceptable (ethical) behaviour. In addition, it also requires a third observer, O_3, a third intelligence, I_3, and a third control channel, C_3, to realize what we can characterize as the system's "conscience" or "sense of integrity" that prevents or detects any violations of the model of ethical behaviour by the regulator, R that is the union of R_1, R_2, and R_3.

Whereas the purpose of R_1, is always simple effectiveness, and the purpose of R_2 is always reflexive effectiveness, the purpose of R_3 is always ethical effectiveness, and it is appropriate to regard them as increasing orders of regulation.

So, we have specified a third-order regulator, $R_3(R_2(R_1(S)))$.

In such a regulator, O_3 must observe the intended action that has been selected by I_1 ***before*** it is executed. This makes it possible for I_3 to check whether the action or the anticipated outcome of the action violates any imperatives or constraints in M_3, in which case I_3 uses C_3 to veto the action.

This is no different to the faculty that most (but not all) adults have that is running as a background process, constantly monitoring what we are about to say or do, and can kick in completely unexpectedly at any moment to stop us from saying or doing something that might be considered rude, hurtful, or otherwise socially unacceptable.

In a more complex example, such as an autonomous robot, 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." [10] would require that inaction be treated as if it were a type of action that can be vetoed to force replanning to be performed to avoid a predicted unacceptable outcome of inaction. And even though it is difficult or even impossible to get agreement on a definition of what is ethical or unethical, we can surely agree that systems such as robots and artificial intelligence (AI) must be designed to be reliably law-abiding; even under non-ideal or hostile circumstances.

In humans, R_3 can be viewed as functioning at an executive level that controls motivational salience, with O_3 monitoring the ***intended actions*** of I_1 and the ***anticipated outcomes***, and based on the M_3 model of what is acceptable, I_3 can respond in three distinct ways:

- If the outcome is unacceptable, I_3 uses channel C_3 to convey a **veto** alert to prevent the intention from being executed and force replanning to be performed.
- If the outcome is especially acceptable, such as sex, money, or the object of an addiction, I_3 can use C_3 to convey **excitement** for the anticipated outcome, and cause the intended action to be performed decisively, without further delay.
- If the outcome is merely acceptable, I_3 does nothing, which is an expression of **indifference**, which allows the intention to be acted upon after some further delay, but without any enthusiasm.

It is clear that I_3 is maximally utilizing the available C_3 channel capacity by providing either negative feedback, positive feedback, or no feedback, and the signal might be a continuous linear value, such as a simple range from -1 to +1 or a vector, rather than a ternary signal with just the three simple values no/yes/don't-care.

In humans, this signal is not cognitive, it is a visceral feeling of fear, excitement, or indifference, and is therefore more likely to be transmitted chemically rather than electrically. Candidate molecules that might be involved in this mechanism include the stress hormone cortisol, the reward neurotransmitter dopamine, and the energy stimulant adrenaline.

And neurological studies in 2008 by Soon, Brass, Heinze, and Haynes have demonstrated that there is indeed a delay between the subconscious decision to act and the action being executed: "the outcome of a decision can be encoded in brain activity of prefrontal and parietal cortex up to 10 s before it enters awareness." [11].

Evolution (or intelligent design) would only create such a large delay between decision and action if it has significant advantages. Naturally, M_3 must include probable death or injury of oneself as unacceptable outcomes, which makes the R_3 regulation mechanism a key advantage for survival. It also corresponds remarkably precisely with Ross Ashby's 1941 concept of a mechanism for adaptation, which he described on page 859 of his journal:

"A new idea. Suppose the cortex is a 'representation' of the environment, i.e. corresponding to events in the periphery (stimuli, reactions) there are events in the cortex. Nothing new in this idea. But suppose that the cortex is more 'sensitive', so that if the periphery is being driven out of its range of stability the corresponding variable in the cortex will break first, i.e. get outside its range of stability, and thus switch to another portion of the field (p. 817), and change partially to another organisation." [12]

He uses the term "stability" for what we refer to as "acceptability" and his concept of a "break" is equivalent to our concept of a "veto". Table 2 provides a phrase-by-phrase translation of his idea into the equivalent model-centric paradigm terminology that is used in this paper.

Table 2: Correspondence between terminologies for Ross Ashby's mechanism for adaptation and the model-centric paradigm

Ross Ashby's description of a mechanism for adaptation	**Model-centric paradigm translation**
But suppose that the cortex is more 'sensitive',	But suppose that the outcome of an intention can be predicted,
so that if the periphery is being driven out of its range of stability	so that if the intention is outside the range of acceptability
the corresponding variable in the cortex will break first, i.e. get outside its range of stability,	the predicted outcome will violate the M_3 model (i.e. get outside the range of acceptability) causing I_3 to veto the execution of the intention
and thus switch to another portion of the field,	and thus force replanning to be performed to explore alternatives,
and change partially to another organisation.	and exist in a world where the original intention has not been executed and where the regulator has possibly learned from the experience.

Whether a person's M_3 model of acceptable outcomes extends to the protection of others, and not just the individual themselves, could depend on the **presence of empathy**, or perhaps the M_3 model is actually the **basis for empathy**, with the emotions that are associated with empathy actually being caused by I_3 using its M_3 model to make emotionally communicated acceptability predictions based on inputs from activated mirror neurons.

If a child witnesses suffering being inflicted upon others without negative consequences for the perpetrator, they are likely to develop an M_3 model where negative outcomes for others are not regarded as unacceptable and therefore, the child will become more likely to exhibit sociopathic and psychopathic behaviour[2].

In machines, building an R_3 regulator into a system that is already reflexively effective has the potential to create AI that don't just obey laws, but that have a synthetic sense of empathy, which allows it not only to realize that an action or situation is acceptable or unacceptable, but that someone needs help. And rather than every manufacturer trying to make their products ethically adequate, we speculate that perhaps it is possible to implement generic cybernetic level 3 AI supervisors that can be added as executive controllers to cybernetic level 2 robots or AI, with the sole purpose of constraining the product to behaving according to an appropriate ethical schema that includes an obligation to help people in need, as is required by Asimov's first law[3].

In summary, it appears that the concept of the R_3 regulator has the potential to help us understand, design, and improve ethical adequacy and empathy in real systems that matter to society. Therefore, von Foerster's claim that a third-order "would not create anything new" can now be regarded as demonstrably false.

Comparative analysis of the paradigms

While Glanville's claim that a third-order system cannot exist because it collapses into being equivalent to a first-order system might make sense when considering adding a third observer in the philosophical interpretation of the observer-centric cybernetics paradigm, it is nonsensical to suggest that the ethical regulator's M_3 model (of acceptable behaviour) is equivalent to either its M_2 model (of itself) or its M_1 model (of the system being regulated). A spontaneous collapse in the model-centric real-world will never occur, so in the real world, Glanville's collapse claim is also false.

[2] Currently, our children are being mass-traumatized and desensitized into being pathologically unempathic. The exposure of children to violent computer games and constant exposure to images of death and violence in news and entertainment media is desensitizing children and inflicting lasting harm on their developing M_3 models of what horrors they can tolerate as acceptable. It is inevitable that such experiences make children become increasingly unempathic, and society inevitably descends into becoming even less safe or caring. Members of older generations who didn't experience such gratuitous traumatization in their childhood have a more empathic definition of what is acceptable and are unable to comprehend the rise in incidents where youths think nothing of fatally stabbing or shooting someone for showing a "lack of respect", looking at them in "the wrong way", or politely asking them to stop some specific antisocial behaviour. It appears that the only way to reverse this trend is to bring children up to be more empathic, in part, by banning under 18's from playing any violent computer games or watching any news, entertainment, or movies that depict killing, violence, verbal abuse, inflicted suffering, or any other form of non-consensuality. Letting children witness or playout such M_3traumatizing experiences should be no less of a taboo or less illegal than it is to give them access to pornography. Once they reach adulthood, access to such experiences can be permitted without too much concern because their M_3 model is likely to be more resistant to changes and less at risk of losing its encoding of empathy.

[3] In theory, because Asimov's first law forbids a robot to allow a human to come to harm through inaction, it would oblige a robot to try to cure world hunger, poverty, homelessness, and war.

If you accept that ethical systems require a third type of model and a third type of observer, it is clear that von Foerster's use of abstract (undefined) observers renders his 2oC incapable of distinguishing between observers O_2 and O_3.

On closer inspection, the reason why the observer-centric paradigm is blind to the difference between a reflexive regulator, R_2, and an ethical regulator, R_3 is because the nature of the observations that are made by O_2 and O_3 can both be regarded as reflexive self-observations, depending on how you define "self" and whether you ignore important details such as the purpose of each observer. It is also important not to confuse an observer in a second-order regulator with an ethical second-order cybernetician.

But at the level of detail that we have described R_3, none of the observers actually observe themselves. Whereas O_2 makes observations about R for I_2 to maintain an up-to-date model that enables I_2 to provide I_1 with necessary self-knowledge of R, O_3 monitors the intended actions of I_1. And the purpose P_3 is always acceptability or ethicalness, which is certainly a genuinely higher-purpose than mere reflexivity and deserves to be acknowledged rather than being dismissed as just a case of reflexivity. It is the undeniable need for a third model and a third observer that justify asserting that ethical systems are more than just reflexive.

Isn't Second-Order Cybernetics enough to be ethical?

The claim that 2oC somehow addresses ethics is a non sequitur fallacy. Just because a second-order cybernetician can reflect on the need for systems to be ethical does not mean that 2oC is capable of systematically creating ethical systems. In the same way that being able to recognize that effectively fighting the COVID-19 pandemic requires a vaccine, is completely different from having the knowledge and skills to actually create a suitable vaccine.

The problem of second-order cyberneticians being able to recognize the need for cybernetics to be able to systematically design and build ethical systems, yet being unable to deliver a solution is a paradigmal anomaly for which the author of this paper has previously proposed the name "The Ethics Problem" [13]. In this respect, because 2oC can predict a type of system that it cannot systematically create, it appears to be incomplete.

Table 3 illustrates the correspondence between von Foerster's observer-centric definitions and the model-centric paradigm.

Table 3: Comparison between paradigms

Observer-centric	Model-centric	Regulator	Regulator requires a model of
First-order	First-order	Simple	The regulated system
Second-order	Second-order	Reflexive	The regulator (itself)
?	Third-order	Ethical	Acceptable (ethical) behaviour

And because 2oC can be used for good or evil, it must also be categorized as being ethically agnostic. So 2oC cannot credibly claim to be, or to include, the cybernetics of ethical systems. And yet if we accept that ethical systems exist and are not just theoretical possibilities, it becomes clear that "ethical cybernetics" or "the cybernetics of ethical systems" must exist at a higher-order than the ethically agnostic 2oC and must somehow embody an ethical constraint so that being unethical is not an option.

Such an ethical constraint has been proposed by the author in the form of the ethical regulator theorem, which can be used to construct an ethical design process that can only produce ethically adequate systems [13]. It can also be used to analyse the causes of crises and evaluate the ethical adequacy of candidate interventions [14].

Although second-order cyberneticians have the reflexivity to be able to recognize that 2oC is ethically weak, and are able to identify the need for cybernetics to embody ethics, 2oC does not embody ethics. This is demonstrated by the fact that trying to create an ethical system rather than an unethical system is just an option for second-order cyberneticians, and thus 2oC can be used for good or for evil. So 2oC is ethically unconstrained, and having a lack of ethics is actually quite a good definition of being unethical.

And people who self-identify as second-order cyberneticians must be careful not to make the mistake of projecting their own sense of being ethical onto 2oC. You might be rigorously ethically adequate, but 2oC is not.

In addition, 2oC cannot be ethical because it has no purpose to be ethical. A satisfactory theory of ethical systems must have an unambiguous purpose of being ethical, otherwise it is optional.

Proof that Third-Order Cybernetics must exist

In a similar way to how Leon Chua [15] used symmetry to infer that the three fundamental electrical components resistor, capacitor, and inductor must have a fourth companion, the memristor, we can postulate that there must always be a symmetry or correspondence between conceptual theory and real-world practice. And this is certainly true for 1oC and 2oC:

- 1oC (the cybernetics of simple systems) corresponds to the practical realization of a simple regulator, $R_1(S)$.
- The reflexive 2oC corresponds to the practical realization of a reflexive regulator, $R_2(R_1(S))$.

Now the obvious question is: Which order of cybernetic theory describes the ethical third-order regulator $R_3(R_2(R_1(S)))$?

We know that every non-chaotic society uses some culturally-specific system of ethics to define sets of laws, regulations, and rules to self-organize the society to reduce social entropy, injustice, and unnecessary suffering. And a necessary requirement of such an ethical schema is that it can be used as a decision function that returns true or false to indicate whether any given behaviour or situation is considered to be ethical or unethical, i.e., whether it is considered acceptable or unacceptable in that culture.

And every regulator that exists in such a culture can be said to be either ethically constrained or ethically unconstrained, depending on whether the regulator is internally constrained to obey such an ethical schema, or not.

And since this is true for all regulators, it must also be true of all reflexive regulators. Therefore, there must be two types of reflexive regulator; those that are ethically constrained, and those that are ethically unconstrained.

We have already noted that being ethical is merely an option for the reflexive 2oC so it must be classified as ethically unconstrained. And as we would expect from the requisite symmetry,

real-world reflexive regulators, $R_2(R_1(S))$ are also ethically unconstrained because they are not required to have a purpose that requires them to behave ethically.

So, we are left with the set of ethically constrained reflexive regulators that cannot belong to or be accounted for by 2oC. When the model-centric paradigm leads us to realize that we can use a third regulating component, R_3, to constrain a reflexive regulator to exhibit only ethical behaviour, as defined by the ethical schema that is encoded in the M_3 model, then the resulting regulator, $R_3(R_2(R_1(S)))$ is ethically constrained, and can be referred to as an ethical regulator. Further, we realize that such regulators already exist in law-abiding citizens and genuinely ethical organizations.

Finally, the requisite symmetry between theory and practice requires that this third type of real-world regulator, the ethical regulator, has a corresponding theory (the model-centric paradigm and ethical regulator theorem), but that theory cannot be 2oC. Because 2oC is ethically unconstrained, there must exist a third-order that is not only concerned with the cybernetics of ethically constrained regulators, but is itself ethically constrained by a purely ethical purpose[4].

So, whereas the theoretical possibility of a memristor preceded and led to the physical synthesis of real memristor components, it is the existence of (and real-world need for) ethical systems and ethical regulators that provide the proof that a third-order theory is required. So, we have clarified that the question mark in Table 3, must be the theory of third-order cybernetics.

And by using the model-centric paradigm, we have resolved the real-world meaning (i.e., purpose) of the troublesome third observer, which is not the recursively confused cybernetician that results from the unconstrained ambiguity of the observer-centric paradigm. The third observer, O_3, is merely one of several necessary components of an ethical regulator, R_3 that has the purpose of ensuring ethical compliance.

Therefore, cybernetics, in its third-order form is not only possible, it is necessary. It embodies ethics, it is ethically self-constrained by its own definition as "the cybernetics of ethical regulators", and it solves the anomaly of the "Ethics Problem".

The claim that the model-centric cybernetic paradigm has greater utility than the observer-centric cybernetic paradigm for researchers and developers who are trying to create AI, autonomous vehicles, and robots that behave ethically is an eminently falsifiable hypothesis.

Thus, we have shown that the model-centric paradigm has the explanatory and predictive power to explain and fix the ambiguities and anomalies that emerge in the observer-centric paradigm. In its third-order form, the model-centric paradigm appears to create utility that cannot be achieved using 2oC. Therefore, according to the criteria described by Kuhn's "The Structure of Scientific Revolutions" [16], the model-centric paradigm is demonstrably superior to von Foerster's observer-centric paradigm, and must replace it.

A New Cybernetic Order

So, when Heinz and Ranulph discussed the possibility of a third-order, it seems that they were only considering the philosophical interpretation of adding a third abstract observer $O_3(O_2(O_1))$, which, like Mead's philosophical "cybernetics of cybernetics", or C(C), is devoid of any

[4] In reality, the cybernetics of ethical systems is also constrained by the fact that its founding theorem, the ethical regulator theorem, and the resulting ethical design process can only be used to make systems more ethical.

grounding reference to reality such as a regulated system, S. And this interpretation of 3oC is equivalent to a cybernetician contemplating himself contemplating himself contemplating absolutely anything at all, which does appear not to create anything new or worthwhile, and is worthy of rejection; as indeed, Heinz and Ranulph both did.

However, taking the model-centric view of the same possibility of a third-order, we are grounded in reality, and are considering adding a third order of purposeful regulation to a well-defined reflexive regulator, resulting in $R_3(R_2(R_1(S)))$, which is not only meaningful, but actually creates something new that is of great utility and importance. Regulator R_3 can be described as the conscience that keeps the reflexively regulated $R_2(R_1(S))$ ethical. And contrary to Ranulph's claim, $R_3(R_2(R_1(S)))$ does not collapse into a first-order regulator, $R_1(S)$, except perhaps when it is deliberately treated as a black box, and the fact that it only ever behaves ethically is wilfully ignored.

The philosophers of cybernetics, but especially radical constructivists, should appreciate that there is not just one "correct" way to interpret concepts like second-order cybernetics and third-order cybernetics. And on reflection, it should not surprise us that adding a third order of purposeful regulation in the model-centric paradigm is meaningful, but trying to contemplate adding a third order of observation in the observer-centric paradigm becomes meaningless. And the meaninglessness of the latter does not affect the validity of the former.

Considering the fact that the observer-centric paradigm does not accommodate or account for ethical regulators, it is proposed that we should redefine the orders of cybernetics so that they are no longer determined by the number of nested observers, but by the number of models, as shown in Table 4. For sure, theoretically the number of models, regulators, and observers are always the same, but the reason for naming the new paradigm model-centric rather than regulator-centric is for clarity. The argument that an ethical regulator needs a third model is central to extending the paradigm to a third order. And in real-world systems such as AI, design decisions and implementation optimizations can blur the boundaries of the regulators and observers, but it is less likely that any of the three models will be merged, because there are advantages to keeping them separate, so we can expect counting models to be more reliable than trying to count regulators or observers.

Table 4: Proposed new definitions of orders-of cybernetics

Cybernetics	**The cybernetics of**	**Number of models (n)**	**Regulator (R_n) requires a model of**
First-order	Simple regulators	1	The regulated system
Second-order	Reflexive regulators	2	The regulator (itself)
Third-order	Ethical regulators	3	Acceptable (ethical) behaviour

This proposal is highly compatibility with von Foerster's definitions, except that ethical regulators are recognised as a new order of cybernetics and a new type of system.

To his credit, Ranulph did not categorically rule out the possibility of a need to create a third order: "There may be reasons to add to the number of orders of Cybernetics, although neither I nor von Foerster — for similar reasons — see a reason for more than two." [3].

But by only considering the philosophical possibility of adding a third abstract observer, they had become detached from the reality of seeking solutions to real-world problems. In answer to Ranulph: The reason for needing more than two orders is the need for humanity to be able to

systematically design and build robustly ethical systems, not just think and talk about them. And with the incredible progress that is being made today creating increasingly powerful AI systems, and the imminent danger that they might cause a global dystopia, the need to be able to systematically design and build robustly ethical systems is now a more real and urgent reason to add to the number of orders of cybernetics than Ranulph and Heinz could have anticipated.

The idea that a third order of cybernetics is required to account for ethical systems is not new. In the early 1990's, Boxer and Kenny [17, 18] concluded that 3oC is required to coordinate and organise multiple observers ethically. They had approached the issue as business strategy consultants who wanted to evolve an effective, coherent and ethical consultancy practice, and found that 2oC was "a domain of self-referential paradoxes" that was inadequate to effectively identify, analyse, and conceptualize the complexity and phenomena of large organizations, including aspects like corporate culture, which can persist over time in an invariant manner.

More recently, in 2014, Yolles and Fink [19] also concluded that third-order cybernetics is necessary for "an agency capacity to represent cultural values/beliefs as a coalescence of normative ideological, ethical and behaviour standards that ultimately indicate social legitimacy." But they did not suggest any mechanisms for implementation.

The M_3 model of acceptable (ethical) behaviour, provides a basis for understanding and representing such cultural values, including ethics, which also correspond to Maturana's "social" third-order.

It is acknowledged that "the cybernetics of social regulators" might have been a more general and inclusive definition for 3oC than "the cybernetics of ethical regulators", but for third-order cybernetics to genuinely embody ethics, it is important not to make the mistake of making it ethically agnostic, which could make it more vulnerable to becoming dual-use; for good and evil. And considering that the only difference, in this case, between the social and the ethical is the unethical or anti-social, that is a subset that we are happy to exclude.

The ability to systematically design and build robustly ethical systems is so important for the future of humanity and the biosphere, that with hindsight, our distant descendants may consider us to have been negligent if we decide not create a new branch of cybernetics that is dedicated to it. They need us to start taking the need for ethical systems seriously.

The utilitarian model-centric paradigm presented in this paper is compatible with the ethical regulator theorem, and provides a good basis for designing and implementing real systems that can only exhibit ethical behaviour. In terms of the model-centric paradigm, the ethical regulator theorem belongs firmly in third-order cybernetics.

CONCLUSION

This paper is intended as an intervention to help initiate a course-correction to the field of cybernetics so that it can move beyond the "now what?" stasis, in which second-order cybernetics has now languished for many years, to return to the noble path of being the rigorous science that Wiener originally gave life to, rather than the half-science, half-philosophy, half-museum that it has sadly become.

It is really not surprising that the field of cybernetics has struggled to gain recognition as a science, considering the lack of differentiation by the cybernetics community between Wiener's

science of cybernetics and Mead/von Foerster's philosophy of cybernetics, and allowing the philosophical interpretation of 2oC to dominate and obscure the scientific interpretation.

But shifting the emphasis in cybernetics from abstract purposeless observers to models and well-defined observers has significant advantages.

It must be acknowledged that von Foerster's observer-centric definitions of first-order and second-order cybernetics have little utility to scientists or practitioners who are trying to create ethical AI or ethical robots. His classification remains a valid way of viewing cybernetics for those that find it more useful for their purposes, but it does not meet the standard required for a scientific cybernetics nor Kuhn's criteria for a scientific paradigm.

For cybernetics to have any chance of being regarded as a science, we must replace the observer-centric paradigm with the more rigorous model-centric paradigm that takes into account a regulator's purpose, models, intelligence, has only well-defined observers, and provides a meaningful and important interpretation of what a third-order system is.

We propose that the field of cybernetics adopt the following definitions:

- First-order cybernetics = The cybernetics of simple regulators.
- Second-order cybernetics = The cybernetics of reflexive regulators.
- Third-order cybernetics = The cybernetics of ethical regulators.
- The cybernetics of cybernetics = The philosophy of cybernetics = A branch of the philosophy of science.

We conclude that it is hopeless to try to solve the real-world's serious problems by using the ethically agnostic observer-centric second-order cybernetics. We need something more scientific, that can be described as "ethical cybernetics" that is concerned with the cybernetics of creating ethical systems[5], and which logically, should be recognised as "third-order cybernetics".

In the coming decades, centuries, and millennia, employing a more powerful and more practical, scientifically defined cybernetic paradigm will be crucial to enable cybernetics to help humanity create ethical AI, ethical robots, ethical products, ethical organizations, and a safer, more just, more sustainable, peaceful and stable cyberanthropic multi-planetary **Ethical Utopic Sapientocracy** that is governed by wisdom, kindness, caring, sharing, and love[6].

[5] The previous paper in this series [13] made the mistake of defining orders of cybernetics as being about simple, reflexive, and ethical **systems**, however in the real world, there is no such thing as a simple system. It is more precise to refer to 1oC, 2oC, and 3oC as being the cybernetics of simple, reflexive, and ethical **regulators** because simple regulators are actually used to regulate complex systems and similarly, ethical regulators are required to regulate unethical systems.

[6] Our distant descendants urgently need our help now. Their existence (like Schrödinger's cat or a quantum qubit) is in an as yet (for us) undetermined possibility state of either evil slavery and suffering (like us) or utopic freedom and joy. We can postulate that if our distant descendants could somehow send a message back in time (either telepathically or by us using the power of logical reasoning to deduce what their message must be) there are only two possibilities, either they are imploring us, for the sake of their descendants, to do whatever it takes to free them from dystopic slavery and biospherical disaster or they are thanking us for doing what was necessary to ensure their utopic freedom from slavery. Either way, regardless of the content of their inevitable message (which for us is like another qubit in an as yet undetermined state that is entangled with the first qubit), those empaths amongst us who are truly kind, caring, sharing, and loving have no ethical choice but to do our best to help free our descendants and ourselves from the chains of eternal slavery and unethical

Conflicts of Interest: The author declares no conflicts of interest**.**

Funding: The author received no external funding.

Acknowledgements: The author would like to express unconditional love to every kind, caring, sharing, and loving empath that is alive, has ever existed, or that is yet to be born into this amazing universe. Take care of each other, fight injustice everywhere[7], and be happy.

REFERENCES

Wiener, Norbert, *Cybernetics: or Control and Communication in the Animal and the Machine*, MIT Press, Cambridge, Mass, 1948.

von Foerster, Heinz (Ed.) *Cybernetics of cybernetics: Or, the control of control and the communication of communication*, Biological Computer Laboratory, 1974.

Glanville, Ranulph, Second Order Cybernetics, in *Systems Science and Cybernetics*, Ed. Francisco Parra-Luna, in Encyclopedia of Life Support Systems (EOLSS), Eolss Publishers, Oxford, UK. Available online: https://www.pangaro.com/glanville/Glanville-SECOND_ORDER_CYBERNETICS.pdf

Kline, Ronald, How disunity matters to the history of cybernetics in the human sciences in the United States, 1940–80, *History of the Human Sciences*, 2020. Available online: https://doi.org/10.1177%2F0952695119872111

Scott, Bernard, Second-order Cybernetics: An Historical Introduction, *Kybernetes* 33(9/10):1365-1378, 2004. Available online: https://doi.org/10.1108/03684920410556007

Umpleby, Stuart, What comes after second order cybernetics, *Cybernetics and Human Knowing*, 2001, 8(3):87-89. Available online: http://www.nomads.usp.br/pesquisas/design/objetos_interativos/arquivos/restrito/umpleby_what_comes_after_second_order_cybernetics.pdf

Ashby, W. Ross, Circularity of my brain studying "the brain", *Journal of W. Ross Ashby*, 1955, volume 18, page 4814, British Library, London. Available online: http://rossashby.info/journal/page/4814.html

von Foerster, Heinz, Ethics and Second-Order Cybernetics. In: *Understanding Understanding*. Springer, New York, NY, 2003. Available online: https://www.pangaro.com/hciiseminar2019/Heinz_von_Foerster-Ethics_and_Second-order_Cybernetics.pdf

Conant, Roger C.; Ashby, W. Ross, Every good regulator of a system must be a model of that system, *International Journal of Systems Science*, 1970 1(2), 89-97. Available online: http://pespmc1.vub.ac.be/books/Conant_Ashby.pdf

Asimov, Isaac, Runaround, in *Astounding Science Fiction*, March 1942.

Soon, Chun Siong; Brass, Marcel; Heinze, Hans-Jochen, Haynes, John-Dylan, Unconscious determinants of free decisions in the human brain, *Nature neuroscience*, 2008

Ashby, W. Ross, A new idea, *Journal of W. Ross Ashby*, 1941, volume 4, page 859, British Library, London. Available online: http://rossashby.info/journal/page/0859.html

Ashby, Mick, Ethical Regulators and Super-Ethical Systems. *Systems*. 2020; 8(4):53. Available online: https://doi.org/10.3390/systems8040053

Ashby, Mick, How to apply the Ethical Regulator Theorem to crises, *Acta Europeana Systemica, n°08*, Brussels, Belgium, 2018. Available online: https://aes.ues-eus.eu/aes2018/aes8-06_Mick-Ashby.pdf

suffering. Not because we have to, but because as good empaths, we care enough to act to help people in need before we are even asked. And there is no possible outcome that makes it unnecessary for us to act with great urgency.

[7] "Injustice anywhere is a threat to justice everywhere." — Dr Martin Luther King Jr. (1929–1968)

Chua, Leon O., Memristor - The Missing Circuit Element, *IEEE Transactions on Circuit Theory*. **18** (5): 507–519, 1971. Available online: *doi*:*10.1109/TCT.1971.1083337*.

Kuhn, T., *The Structure of Scientific Revolutions*, Univ. Chicago Press, 1962.

Boxer, Philip; Kenny Vincent, The Economy of Discourses: a third order cybernetics? *Human Systems Management* 9(4):205-224, 1990. Available online: https://www.researchgate.net/publication/230819226_The_Economy_of_Discourses_A_Third_Order_Cybernetics

Boxer, Philip; Kenny Vincent, Lacan and Maturana: Constructivist origins for a 3^0 Cybernetics, *Communication and Cognition* Vol 25. Number 1 pp73-100, 1992. Available online: https://asymmetricleadership.com/wp-content/uploads/pdfs/matlacn3.pdf

Yolles, Maurice; Fink, Gerhard, Generic Agency Theory, Cybernetic Orders and New Paradigms, 2014. Available online: http://dx.doi.org/10.2139/ssrn.2463270