# A Primer For Conant & Ashby's
# "Good-Regulator Theorem"

*by Daniel L. Scholten*

# Table of Contents

## *Introduction*

Unless you have sufficient confidence in your math skills and especially a basic familiarity with information theory, it can seem a daunting task to understand the proof given by Roger C. Conant and W. Ross Ashby (henceforth C&A) of their theorem that establishes that "every good regulator of a system must be a model of that system"[1]. The purpose of the present paper is to train the reader to accomplish this task.

As to why any reader might want to accomplish this task I will only say here that Humanity's increasing dependence on its already huge, still growing, and increasingly complicated system of models and representations strongly suggests that *every* civilized person really should have some sort of enriched and high-level understanding of this fundamental result from the System Sciences. It is my belief that this theorem ought to be included in the standard science curriculum along side such other basics as the germ-theory of disease, the sun-centered view of the solar system and the meaning of symbolic scribbles such as "2+2=4". I have elsewhere explored this topic in greater detail and in this essay I will focus somewhat narrowly on the theorem itself[2].

In the pages that follow I will attempt to provide a self-contained exposition of the proof of this "Good-Regulator Theorem" in a way that requires a minimum of prerequisites so that any literate adult with sufficient motivation and some prior experience with very basic probability theory, the logarithm function, and perhaps a very little bit of calculus[3] might follow their argument. Although this will substantially lengthen the argument (44 pages to explain what amounts to a single page in the original article), the hope is that the much longer argument will be easier to understand.

Please notice that I said that I will "attempt" to provide this. Whether I succeed will have to be determined by readers like you, and if you should determine that I have fallen short of this goal, you are invited to send me your suggestions for improvement so that future versions of this essay will come closer to it[4].

Let's begin with the authors' main objective:

---

[1] Roger C. Conant and W. Ross Ashby, "Every Good Regulator of a System Must be a Model of that System," *International Journal of Systems Science*, 1970, vol 1., No. 2, 89-97.

[2] For an accessible, non-specialist treatment of the theorem, see *The Three Amibos Good-Regulator Tutorial.* A somewhat more advanced technical analysis can be found in "Every Good Key Must Be A Model Of The Lock It Opens". Both of these resources are available free of charge on the "Education Materials" page at www.goodregulatorproject.org .

[3] The brief passage involving Calculus can be skimmed with little loss to overall understanding.

[4] Please forward your comments to me at dlscholten@goodregulatorproject.org.

## *What the theorem claims*

We will begin by taking a closer look at what C&A actually proved by their theorem. As the title of their paper proclaims: "Every good regulator of a system must be a model of that system". This assertion uses the terms *regulator*, *system* and *model*, and because each of these can be interpreted in various ways we are going to first clarify what they mean in the context of the Good-Regulator Theorem.

First of all, the terms *system* and *regulator* are being used here to refer to dynamic entities, meaning that they can exhibit various state-changes. We will refer to these state-changes somewhat colloquially as "behaviors", so that even a randomly changing system, such as a weather system, will be described as exhibiting behaviors. Furthermore, the regulator in question is such that its behaviors can be *goal-directed*, which is to say that it can execute its behaviors in the service of some preferred state-of-affairs. Although this does not imply that the regulator must therefore be a human being (a mechanical device such as a Watt Governor can be legitimately described as "goal-directed") it just happens to be the case that human beings make great examples of such regulators. Now, it might also be the case that the system in question is goal-directed, but in the current context this is not a necessary attribute of the system. The system *might* be a goal-directed human being, but it might also be no more goal-directed than the weather.

Another point to recognize is that the system and the regulator are *interacting*. As it concerns the current context, what this means specifically is that the regulator is attempting to attain a goal and the system keeps doing things that make it difficult for the regulator to accomplish this. In this sense, then, the system is really a system of obstacles and it is the regulator's job to handle or *respond* to those obstacles in such a way that the goal is achieved. Also, the Good-Regulator Theorem makes specific reference to something called a "good regulator". In this context, the word *good* means quite specifically that the regulator in question is both optimal and maximally simple. It is *optimal* in the sense that it does the best possible job of achieving its goal under the given circumstances, and it is maximally *simple* in the sense that it does this best-possible-job with the least possible amount of effort or expense.

Next, the term *model* is being used to describe a representational relationship that can exist between some given object – the so-called "model" – and some other object – the thing being modeled. Here we have to be somewhat careful about what this actually means. The term *model* has a colloquial interpretation that tends to imply a *visual* resemblance between the model and what it represents, but this sort of definition is too narrow for our purposes. It would exclude, for example, the type of representational activity that goes on, say, whenever we make a grocery list. Such a list can hardly be said to have any sort of visual resemblance to the actual items represented on the list, and yet a grocery list is clearly a representation of those items. Or consider the type of modeling that is used during a Monday lunch-hour recap of

Sunday's football game: the pepper shaker represents the quarterback, the ketchup bottle represents a lineman, etc., and yet none of these items bears any sort of visual resemblance to an actual football player. Although our definition of *model* will certainly allow for the *possibility* of such visual resemblance, it will not be a requirement. Instead, the definition we will use is grounded in the mathematical idea of a *mapping* (a.k.a. *Function*)[5]. The essence of such a mapping is that it somehow associates all of the various component "bits and pieces" of the thing-modeled to the component "bits and pieces" of the model. In the current context, the model is going to be the regulator, the thing-modeled will be the system and the component "bits and pieces" will be their respective behaviors.

A specific example will make all of this clearer while introducing some of the mathematical symbolism we will need to prove the theorem. First, let's suppose we are hoping to regulate some system, call it $S$[6], which can exhibit, say, six distinct and mutually exclusive behaviors. Note that the requirement here that the behaviors be mutually exclusive simply means that if it seems that the system can do two or more behaviors at once – perhaps sing a song while standing on its head – then such composite behaviors need to be treated as separate behaviors. Thus, a behavior such as singing while standing on it head would be considered distinct from either merely singing or merely standing on its head. Using some simple set-builder notation we can represent the complete behavioral repertoire of this system as $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$. Note also that $S$ is understood to be the *complete* behavioral repertoire of the system meaning that it contains every possible distinct and mutually exclusive behavior that the system could ever execute, even if one of those behaviors is some sort of null behavior such as "sit and do nothing". If we can recognize it as something that the system can do it, then it needs to be represented in $S$ as well. Furthermore, let's suppose that the only thing we really understand about the inner workings of this system is the relative frequencies with which it is executing these behaviors. That is, suppose we have a probability distribution of the form $p(S) = \{p(s_1), p(s_2), p(s_3), p(s_4), p(s_5), p(s_6)\}$ where $p(s_j)$ is the probability that $S$ executes behavior $s_j$. Although for now we won't really do much with this probability distribution, when we get to the actual proof of the theorem we are going to need it and so I want to at least introduce it here.

[5] http://en.wikipedia.org/wiki/Map_(mathematics)

[6] In their original paper, Conant and Ashby emphasized the distinction between a system as a thing and the set comprised of that thing's possible behaviors. They did this by using the symbol "S" (without italics) to represent the thing and "*S*" (with italics) to represent that thing's behavioral repertoire, i.e. the set comprised of every behavior that S can perform. They did something similar with the regulator, using "R" and "*R*" to distinguish between the regulator as a thing and its behavioral repertoire, respectively. Although I agree that this is an important distinction to maintain, I will not use two different font styles to maintain it. Instead, I will rely on the reader's good judgment and ability to recognize the distinction from contextual clues. Thus, I will write *S* or *R* (with italics) to sometimes represent the thing (system or regulator, respectively) and sometimes to represent that thing's behavioral repertoire.

Next, let's suppose that off to the side some where we have built a device that we are hoping to use to regulate that system – a *candidate regulator*, although to streamline our discussion we will refer to this device simply as a *regulator* regardless of whether it is actually regulating the system. Also, let's suppose that we have built this device to exhibit just four distinct and mutually exclusive behaviors and – similarly to what we did with the system – let's call this regulator $R$ and use $R = \{r_1, r_2, r_3, r_4\}$ to represent the complete behavioral repertoire of this regulator.

Now, we have built this regulator as an isolated entity; that is, we can set it running and it will start performing its various behaviors in some sort of sequence, but because it is currently a *separate* entity its behaviors will not interact in any way with the behaviors of $S$. Later on we will consider how to set this regulator up so that it does interact with the system and in fact *regulates* that system, but for now we are only going to consider how it might be set up to represent or *model* that system.

As mentioned above, the sort of modeling relationship we are going to use requires only that the component "bits and pieces" of the system be associated in some way with the component "bits and pieces" of the model, where in our current context these "bits and pieces" are understood to be the behaviors of the system and regulator respectively. Now, having set aside for the time being our hope of using our regulator-device to actually *regulate* the system, and given only our desire to use it as a *model* of that system, there are actually lots and lots of ways this might be done. Consider the following example:

$$h\downarrow \begin{array}{cccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ \hline r_3 & r_1 & r_3 & r_2 & r_3 & r_1 \end{array}$$

The diagram shown above is simply a look-up table that displays the essential details of what is meant by the statement "$R$ is a model of $S$". Reading from the table, we see that the statement "$R$ is a model of $S$" simply means that:

- Whenever $S$ does $s_1$ then $R$ *always* does $r_3$;
- Whenever $S$ does $s_2$ then $R$ *always* does $r_1$;
- Whenever $S$ does $s_3$ then $R$ *always* does $r_3$;
- Whenever $S$ does $s_4$ then $R$ *always* does $r_2$;
- Whenever $S$ does $s_5$ then $R$ *always* does $r_3$; and
- Whenever $S$ does $s_6$ then $R$ *always* does $r_1$.

We can observe two other points about this table. First of all, the table has a name – "$h$", and second, there is also an arrow that helps us distinguish between the model

and the thing-modeled. Notice that this arrow points downward *from* the system side of the table *toward* the regulator side of the table. This is meant to show that in this case it is the regulator that is the model and that it is the system that is the thing being modeled. The standard mathematical shorthand used to symbolize this sort of relationship between $h$, $S$ and $R$ is just "$h : S \rightarrow R$", which is read "$h$ is a mapping from the set $S$ to the set $R$."

The above is just one somewhat arbitrary example of how our regulator could be used to represent or model the system. Here are a few others:

$$f \downarrow \begin{array}{c|cccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ \hline & r_4 & r_4 & r_1 & r_2 & r_2 & r_2 \end{array}$$

$$m \downarrow \begin{array}{c|cccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ \hline & r_2 & r_3 & r_4 & r_1 & r_2 & r_3 \end{array}$$

$$q \downarrow \begin{array}{c|cccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ \hline & r_3 & r_3 & r_3 & r_3 & r_3 & r_3 \end{array}$$

Remember that we are not yet saying anything about whether the regulator is actually regulating the system. Maybe it is and maybe it isn't. We will get to that shortly, but at this point we are only examining this idea of using the regulator as a model of the system. In order to establish such a representational relationship between the regulator and the system we first need to map the system's behaviors to the regulator's behaviors, and there are lots and lots of ways this might be done. The above mappings ($h$, $f$, $m$ and $q$) are just four of these.

Another important point to recognize about these sorts of mappings is that in each case *all* of the elements in the set $S$ are associated to an element in the set $R$, but the opposite is not necessarily true. In the last example shown ($q : S \rightarrow R$) all of the system's behaviors are being represented by just one of the regulator's behaviors and the regulator's remaining behaviors do no real "representational work" at all. This situation corresponds to the type of modeling being done, say, when we use a pepper shaker as a model of a quarterback. In that case, all of the complex "bits and pieces" of the actual quarterback are mapped on to a single simple attribute of the pepper shaker – the fact that it is a solid, discrete object that can stand – and all of the pepper shaker's other "bits and pieces" (the pepper, the screw-on lid, the glass body, etc.) do no actual "representational work". On the other hand, the situation illustrated by $m : S \rightarrow R$ – in which all of the regulator's behaviors have been used but because $R$ has fewer behaviors than $S$, some of these do more "representational work" than the others – this type of situation corresponds to the use, say, of a toy car as a model of a real car. In that case, all of the obvious component "bits and pieces" of the toy car –

the wheels, windshield, bumpers, etc. – have been associated with analogous "bits and pieces" of the real car, but some of these have been used more than once. For example, the simple slab of plastic that runs along the bottom of the toy car is used to represent all of the complex "bits and pieces" that are beneath the real car – the muffler, chassis, break lines, etc.

To summarize the above analysis of the term *model*: whenever we have two sets – call them $A$ and $B$ – and some mapping – call it $t$ – where this mapping is *from* the elements in $A$ *to* the elements in $B$, then we will symbolize this relationship between the three of them by writing $t : A \to B$, and we will also say that $B$ is a *model* of $A$. Within the context of our system and regulator, a more colloquial way to paraphrase all of this is to say that "*R* is a model of *S* in the sense that *R*'s behaviors are just *S*'s behaviors *as seen through* some mapping".

Equipped with this much more precise vocabulary, we are now ready to examine what the C&A theorem claims which is that whenever a given regulator is acting as a so-called "good-regulator" of some given system, then it must also be true that the regulator is a model of the system in the sense that the regulator's behaviors are just the system's behaviors as seen through some mapping. Another way to say this last part is that whenever the system executes some given behavior, the regulator always responds *in exactly the same way*. Conversely, to the extent that a regulator *varies* its responses to any given system behavior (and thus ceases to be a model of the system), it must also be the case the the regulator is either not doing as well as it might, or else it is doing so in an unnecessarily complicated fashion.

The author's pack everything we've discussed above into a concise theorem which we can be stated as follows:

*Theorem* : The simplest optimal regulator $R$ of a system $S$ produces behaviors from $R = \left\{ r_1, r_2, \cdots, r_{|R|} \right\}$ which are related to the behaviors in $S = \left\{ s_1, s_2, \cdots s_{|S|} \right\}$ by a mapping $h : S \to R$.[7]

The above is the actual statement that we are working toward proving. (Note: the symbols $|R|$ and $|S|$ are just mathematical shorthand for the number of elements in the sets $R$ and $S$ respectively).

[7]Conant and Ashby's wording is slightly different, but equivalent.

## *Proving the claim*

### The Regulator's Responsiveness to the System

Having clarified what we mean by the terms *system*, *model* and *regulator*, we can now turn our attention to what it means to say that the regulator is *responding* to the system. Furthermore, we will also need some way to represent the outcomes of these responses, but for now we will focus only on $R$'s responsiveness to $S$.

There are various ways we might interpret what it means to say that "$R$ responds to $S$" but C&A wanted to make their result as general as possible and so they chose to use what is known as a *stochastic* (probabilistic) approach. (This is consistent with the probability distribution $p(S)$ introduced earlier for the behaviors in $S$) That is, they chose not to get bogged down in trying to consider all of the various possible particular mechanisms that might be used to make $R$ respond to $S$ and instead they chose to use a method that simply specifies for each possible system behavior a corresponding set of *conditional probabilities* over the entire behavioral repertoire of $R$. This approach is just about as general as we could get and applies to any conceivable way to make $R$ respond to $S$. It can even be used to describe situations in which $R$'s behavior is completely determinate. No matter what the specific technical details might be, such a method will always allow us to make statistical statements of the form, "whenever the system executes $s_j$, there is an $x\%$ chance that the regulator will respond by executing $r_i$". Throughout what follows we will represent such a conditional distribution in one of two ways. Sometimes we will use set-builder notation and represent it as follows:

$$p(R \mid S) = \{ p(r \mid s) : r \in R, s \in S \}$$

And sometimes we will use a table notation and represent it as follows:

| $p(R \mid S)$ | $s_1$ | $s_2$ | $\cdots$ | $s_{\lvert S\rvert}$ |
|---|---|---|---|---|
| $r_1$ | $p(r_1 \mid s_1)$ | $p(r_1 \mid s_2)$ | $\cdots$ | $p(r_1 \mid s_{\lvert S\rvert})$ |
| $r_2$ | $p(r_2 \mid s_1)$ | $p(r_2 \mid s_2)$ | $\cdots$ | $p(r_2 \mid s_{\lvert S\rvert})$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $r_{\lvert R\rvert}$ | $p(r_{\lvert R\rvert} \mid s_1)$ | $p(r_{\lvert R\rvert} \mid s_2)$ | $\cdots$ | $p(r_{\lvert R\rvert} \mid s_{\lvert S\rvert})$ |

In either case, for any regulator behavior, say $r_i \in R$, and any system behavior, say $s_j \in S$, the number $p(r_i \mid s_j)$ is the *conditional* probability that the regulator will respond by executing behavior $r_i$ given the condition that the system executes behavior $s_j$. Note that in dealing with such conditional distributions it should always be kept in mind that each number in the distribution is between 0 and 1 inclusive and that the sum of the numbers in any column must total 1. More formally, we will write:

$$0 \leq p(r_i \mid s_j) \leq 1, \text{ for all } r_i \in R \text{ and all } s_j \in S$$

and

$$\sum_{i=1}^{|R|} p(r_i \mid s_j) = 1, \text{ for all } s_j \in S$$

To return to our current example, since there are an infinite number of real numbers between 0 and 1, then clearly there are an infinite number of ways we might create such a conditional distribution in order to specify out regulator's responsiveness to the system. One (somewhat complicated) example would be the following:

| $p(R \mid S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ |
|---|---|---|---|---|---|---|
| $r_1$ | .20 | .45 | .50 | .85 | 1.0 | .40 |
| $r_2$ | .30 | .19 | .21 | 0.1 | 0 | .31 |
| $r_3$ | 0 | .19 | .19 | 14.9 | 0 | .14 |
| $r_4$ | .50 | .17 | .10 | 0 | 0 | .15 |

What the above distribution tells us, for example, is that whenever the system executes behavior $s_1$ the regulator will never respond by doing behavior $r_3$ (since $p(r_3 \mid s_1) = 0$) but that it will do behavior $r_4$ on roughly half of all such occasions (since $p(r_4 \mid s_1) = .50$), and that it will do behaviors $r_1$ and $r_2$ on roughly 20% and 30% respectively of all such occasions ($p(r_1 \mid s_1) = .20$ and $p(r_2 \mid s_1) = .30$). As another example, the above table tells us that $R$ will execute $r_1$ in response to 85% of the times that the system executes behavior $s_4$, that $R$ will do $r_2$ and $r_3$ on 0.1% and 14.9% of all such occasions respectively, and that it will never do $r_4$ on such occasions. As a third example, the above schedule tells us that the $R$ will always do

$r_1$ in response to the system doing $s_5$ and never do any of its other behaviors in such a situation.

Clearly the above distribution fulfills the two conditions for any such conditional distribution. That is, each number in the distribution is between 0 and 1 inclusive, and if you pick any column and sum the numbers in that column you arrive at a total of 1. The meaning of this latter fact is that the regulator's behavioral repertoire $R = \{r_1, r_2, r_3, r_4\}$ is really a complete inventory of all possible regulator behaviors and since it is a tautology to say that the regulator must always be doing something that it can do then the probability that it does something it can do (given the system has done, say $s_j$) must equal 1. In other words, and since $R$'s behaviors are mutually exclusive,

$$\begin{aligned} & p\{\text{the regulator does something it can do} \mid \text{the system has done } s_j\} \\ = \; & p\{\text{the regulator does } r_1 \text{ or } r_2 \text{ or } r_3 \text{ or } r_4 \mid \text{the system has done } s_j\} \\ = \; & p(r_1 \mid s_j) + p(r_2 \mid s_j) + p(r_3 \mid s_j) + p(r_4 \mid s_j) = 1 \end{aligned}$$

Now, the previously displayed conditional distribution represents just one of the infinite number of ways we might set up the regulator $R$ so that it is responsive to $S$ and most of these infinite ways are rather complicated. But there is a much simpler way we might do this. For example, we might use the following distribution:

| $p(R \mid S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ |
|---|---|---|---|---|---|---|
| $r_1$ | 0 | 1 | 0 | 0 | 0 | 1 |
| $r_2$ | 0 | 0 | 0 | 1 | 0 | 0 |
| $r_3$ | 1 | 0 | 1 | 0 | 1 | 0 |
| $r_4$ | 0 | 0 | 0 | 0 | 0 | 0 |

What this much simpler sort of distribution tells us is that whenever the system does some particular behavior, the regulator's response is *always the same*. (That should sound familiar, for reasons to be explained shortly). For example, given that the system does, say $s_1$, then the probability that the regulator does $r_3$ is 1 (that is, $p(r_3 \mid s_1) = 1$), meaning that it is *certain* that the regulator does $r_3$ whenever the system does $s_1$ and that the regulator will never do any other behavior in its repertoire

when the system does $s_1$ (since $p(r_1 \mid s_1) = p(r_2 \mid s_1) = p(r_4 \mid s_1) = 0$). Of course, the regulator's response might change when the system's behavior changes. For example, according to the above distribution, whenever the system switches to behavior $s_4$ the regulator will always switch to doing $r_2$ but that is the only time the regulator might change its response. As long as the system does the same behavior and whenever it does that same particular behavior, the regulator's response is always the same response to that particular system behavior.

The key thing to notice here is that this is precisely the sort of situation that is described by the term *mapping*. Notice that the result of the above very simple sort of probability distribution is that each one of the system's "bits and pieces" (i.e behaviors) is mapped to exactly one of the regulator's "bits and pieces" – that is, the particular regulator behavior that is performed with probability 1 in response to the given system behavior. In fact, this particular mapping is exactly the same one we considered earlier and which we represented as the following table:

$$h \downarrow \begin{array}{cccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ \hline r_3 & r_1 & r_3 & r_2 & r_3 & r_1 \end{array}$$

You should take a moment to convince yourself that the above table and the previous conditional probability distribution are really just two different ways to specify the exact same mapping $h : S \rightarrow R$. Although the conditional probability distribution is surely the more complicated of the two, we are going to need this extra complication in order to prove the Good-Regulator Theorem.

## The Outcomes Produced by *R*'s Responses to *S*

So, that is how we will represent $R$'s responsiveness to $S$, via some conditional distribution $p(R \mid S)$, which, in the general case will not necessarily specify a mapping from $S$ to $R$, but which at least could do so (and thus make $R$ into a model of $S$). Now it is time to consider *what actually happens* whenever $R$ responds to $S$, that is, the actual outcomes that arise from these responses. The reason these outcomes are important, of course, is that they form the very substance of the regulator's goal.

In order to represent the outcomes that arise whenever $R$ responds to $S$, C&A make use of a different sort of mapping. As with any mapping, this one will also have a name – "$\psi$", the Greek letter, pronounced "sigh" – and involve two sets, but the first of these sets is actually a kind of mixture of the sets $S$ and $R$. This "mixture" is called the *cross-product* of $R$ and $S$, and is symbolized as follows:

The *cross-product* of $R$ and $S$ $= R \times S = \{\langle r,s\rangle : r \in R, s \in S\}$

The elements in this set are known as *ordered pairs* and each ordered pair in that set consists of a single element from the set $R$ along with a single element from the set *S*. For example, if the element from $R$ is $r_i$ and the element from $S$ is $s_j$, then we can symbolize the ordered pair that consists of these two elements as $\langle r_i, s_j \rangle$. The cross-product of R and S, then, is the set that consists of every possible such ordered pair.

That describes the first set involved in the mapping we are calling "$\psi$". The second set is just the set of every possible outcome that could arise from some combination of an $S$ behavior and an $R$ behavior. We're treating the most general case here so we won't be concerned with any particular details of such outcomes, but we will introduce a third set $Z = \{z_1, z_2, \ldots, z_{|Z|}\}$ to represent these possible outcomes. Thus, the purpose of the mapping $\psi$ is to link each particular combination of a regulator behavior and a system behavior, that is, each ordered pair $\langle r,s\rangle \in R \times S$, to a unique result or outcome in the set $Z$. Using the standard shorthand introduced above we can represent this mapping as $\psi : R \times S \to Z$. (Of course, based on our earlier discussion of models, the existence of this mapping means that we can also say that "$Z$ is a model of $R \times S$", but this particular model is not the one we are really concerned with here and so we will just ignore this option. As it concerns our present discussion of the outcomes that arise from the regulator's responses to the system, we are only interested in the actual mapping $\psi : R \times S \to Z$.)

Now, if we happen to know, for example that the particular ordered pair $\langle r_i, s_j \rangle \in R \times S$ maps under $\psi$ to the particular result $z_k \in Z$, then we could represent this particular fact as $\psi(\langle r_i, s_j \rangle) = z_k$, or perhaps the more streamlined $\psi(r_i, s_j) = z_k$, as C&A do, which renders the angle brackets "$\langle\ \rangle$" implicit.

To illustrate all of this with a more concrete example, let's use the sets $R = \{r_1, r_2, r_3, r_4, r_5\}$ and $S = \{s_1, s_2, s_3, s_4, s_5\}$ and for the set of possible results we'll use $Z = \{z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8\}$. The sizes of the sets $R$ and $S$ in this example are a little different from the ones we used earlier, but these differences are superficial and the only reason I've made them different is to illustrate that the differences don't really matter. One thing we should recognize about these sorts of examples is that they are not meant to imply any restrictions on the sizes of the sets $R$, $S$ or $Z$, (represented by $|R|$, $|S|$ and $|Z|$, respectively) either in an absolute sense or relative to each other, and in practice any of these, in fact, may be infinite. Thus, although in this example we have that $|Z| = 8 > 5 = |S| = |R|$ this is really just a matter of haphazard

convenience. In the general case we might have any of the following: $|Z| \leq |S| \leq |R|$, $|Z| \leq |R| \leq |S|$, $|S| \leq |Z| \leq |R|$, $|S| \leq |R| \leq |Z|$, $|R| \leq |S| \leq |Z|$ or $|R| \leq |Z| \leq |S|$.

One point to recognize here is that there are an infinite number of ways that $\psi$ might map elements in $R \times S$ to elements in $Z$. The particular one we will use for our current discussion is arranged in the following table:

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ |
| $r_2$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ | $z_2$ |
| $r_3$ | $z_3$ | $z_7$ | $z_5$ | $z_6$ | $z_7$ |
| $r_4$ | $z_8$ | $z_2$ | $z_2$ | $z_1$ | $z_4$ |
| $r_5$ | $z_3$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ |

The above table, called a "payoff" matrix, is to be read as you would a multiplication table. That is, if you want to know, for example, the particular element of $Z$ to which $\psi$ maps the ordered pair $\langle r_i, s_j \rangle \in R \times S$, then you look in the cell of the table that lies at the intersection of the table's $r_i$ row and the $s_j$ column. Thus, the table specifies that $\psi(r_1, s_1) = z_1$, $\psi(r_4, s_5) = z_4$ and $\psi(r_5, s_3) = z_7$. Note that the table also indicates that $\psi(r_2, s_1) = \psi(r_3, s_4) = \psi(r_5, s_2) = z_6$ which is meant to illustrate the more general possibility that the same outcome (e.g. $z_6$) can be obtained via $\psi$ in response to various different combinations of the elements in $R \times S$ (e.g. for $z_6$ these would be the ordered pairs $\langle r_2, s_1 \rangle$, $\langle r_3, s_4 \rangle$ and $\langle r_5, s_2 \rangle$). A related point to notice is that some of the columns – in particular the columns for $s_2$ and $s_4$ – contain repeated occurrences of the same outcome. For example, the $s_4$ column shows the outcome $z_1$ in both the row for $r_2$ as well as the row for $r_4$. This possible characteristic for such a table will play an important role later on in the argument.

Just as a point of comparison, other possible mappings $\psi : R \times S \rightarrow Z$ for our example sets $R$, $S$ and $Z$ would be any of the following:

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $z_3$ | $z_2$ | $z_6$ | $z_6$ | $z_1$ |
| $r_2$ | $z_3$ | $z_7$ | $z_8$ | $z_1$ | $z_2$ |
| $r_3$ | $z_3$ | $z_8$ | $z_5$ | $z_6$ | $z_7$ |
| $r_4$ | $z_8$ | $z_2$ | $z_4$ | $z_7$ | $z_4$ |
| $r_5$ | $z_5$ | $z_6$ | $z_4$ | $z_8$ | $z_1$ |

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ |
| $r_2$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ |
| $r_3$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ |
| $r_4$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ |
| $r_5$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ | $z_4$ |

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ |
| $r_2$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ | $z_2$ |
| $r_3$ | $z_3$ | $z_4$ | $z_5$ | $z_6$ | $z_7$ |
| $r_4$ | $z_8$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ |
| $r_5$ | $z_5$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ |

And as an illustration of examples involving *different* sets $R$, $S$ and $Z$ consider the following:

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ |
|---|---|---|---|---|---|---|---|
| $r_1$ | $z_{11}$ | $z_4$ | $z_6$ | $z_8$ | $z_3$ | $z_1$ | $z_3$ |
| $r_2$ | $z_1$ | $z_1$ | $z_3$ | $z_5$ | $z_2$ | $z_7$ | $z_5$ |
| $r_3$ | $z_{10}$ | $z_1$ | $z_8$ | $z_5$ | $z_{11}$ | $z_1$ | $z_6$ |

and

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ |
|---|---|---|---|---|
| $r_1$ | $z_1$ | $z_5$ | $z_{11}$ | $z_9$ |
| $r_2$ | $z_{14}$ | $z_2$ | $z_6$ | $z_8$ |
| $r_3$ | $z_{15}$ | $z_1$ | $z_3$ | $z_{13}$ |
| $r_4$ | $z_{10}$ | $z_{15}$ | $z_1$ | $z_4$ |
| $r_5$ | $z_8$ | $z_7$ | $z_8$ | $z_{12}$ |
| $r_6$ | $z_{12}$ | $z_6$ | $z_6$ | $z_9$ |

## Optimal Regulation

Once we have specified a conditional distribution $p(R \mid S)$ and the mapping $\psi : R \times S \to Z$, then we have everything that we need to know both about the regulator's responsiveness to the system as well as what actually happens as a result of that responsiveness. The next question we have to address concerns what it means to say that a regulator is "optimal", or, that it is "doing the best-possible job under the given circumstances". We can also refer to this as "successful regulation".

In order to define what it means to say that a regulator behaves optimally, let's start by considering an actual regulator. For example, suppose we want to purchase a thermostat in order to regulate room temperature. How could we tell if a thermostat is a good thermostat? Clearly, a thermostat is successful when it behaves in such a way as to keep the room temperature constant, or as close to constant as possible, especially when the outside temperatures are fluctuating. Generalizing from this example we might conclude that any regulator can be considered successful if it behaves in such a way as to achieve as much constancy or as little change as possible in the set of outcomes it produces. Another way to say this is that a successful regulator should reduce as much as possible the unpredictability in the set of outcomes it produces.

But hold on. Suppose we purchase a thermostat and discover once it is installed that it does a fantastic job of maintaining a constant room temperature, but that the only

constant room temperature it maintains is 350°F! Should we still consider this regulator successful? Suppose furthermore that it does such a great job of maintaining the room temperature at 350°F that it could do so even if the room were placed on the surface of the planet Mercury where the outside temperatures range from a low of -300°F to a high of 800°F. Now how should we evaluate the success of this thermostat?

On the one hand, 350°F is a lousy room temperature so we might conclude that the regulator in question is doing a lousy job. On the other hand, any device that could maintain a constant room temperature in the face of such extreme outside temperature variation as is found on the surface of Mercury is a pretty amazing device indeed – regardless of what that constant temperature might be. Furthermore, suppose we actually needed to maintain such a constant 350°F temperature under such extreme conditions. That is, suppose we wanted to open, say, a high volume cupcake factory on the surface of Mercury, where we needed to bake cupcakes at 350°F around the clock all year long. In that albeit bizarre situation such a regulator would be just what the doctor ordered. Under such circumstances, such a regulator would clearly be successful.

The point of this example is to illustrate that regulator success can be defined in at least a couple of ways. On the one hand, we might say that a regulator is successful as long as it can minimize changes in the outcomes it produces, regardless of what those outcomes might be. Such a definition would always count as successful such regulators as the above thermostat. On the other hand, we might add the additional requirement that a regulator should be able to receive some user's arbitrary temperature request – 350°F, 25°F, 118°F, etc. – and then adjust itself so as to maintain that requested temperature. Such a definition would exclude the above thermostat in some situations (if we were planning to use it, say, to regulate living room temperature) and it would include it in others (if we wanted to use it to maintain temperature in a cupcake oven on Mercury).

Presumable because the first approach is more general, and because the second approach depends on the somewhat arbitrary requirements of the context in which the regulator is actually used, Conant and Ashby use the first approach. That is, they define regulator success strictly in terms of *stability* – the minimization of the changes in the outcomes produced by the regulator's responses to the system. If it turns out that a given regulator produces a constant (or relatively constant) set of outcomes that happen to be undesirable in one context, well, then that just means we have to find the right context for the regulator, but we will still count it as a good regulator.

Next we need a way to measure the extent to which the regulator is able to achieve such a set of (relatively) constant outcomes. Now, if the outcomes are associated with some numeric variable, such as temperature in the case of a thermostat, then the standard measuring tool would be the statistical variance which is defined as the expected value of the squared differences between the outcomes obtained and the expected value of those outcomes. Formally, letting $X$ represent the numeric variable

associated with the outcome, and letting $E(X)$ represent the expected value of $X$, then the variance of $X$ is defined as,

$$Var(X) = E\left[\left(X - E(X)\right)^2\right]$$

But this approach requires that we have some *numeric* variable we can measure and Conant and Ashby wanted to treat those cases as well as cases in which no such numeric variable was available. In order to accomplish this they use a device from Information Theory called the *Shannon Entropy Function* which is basically a measure of variation that depends only on the probability distribution that governs whatever particular process whose variation we are trying to measure. In the current context the process that concerns us involves the occurrences of the various outcomes in $Z$ that result from combining regulator behaviors with system behaviors and so, letting $p(z_k)$ represent the probability that some particular result $z_k \in Z$ is obtained, the probability distribution of interest is $p(Z) = \left\{p(z_1), p(z_2), \ldots, p\left(z_{|Z|}\right)\right\}$ and the entropy function for the set $Z$ of outcomes is defined as follows:

$$H(Z) \equiv -\sum_{k=1}^{|Z|} p(z_k) \log p(z_k)$$

Thus, C&A define an optimal regulator as one that responds to the system so as to make the entropy $H(Z)$ as small as possible, given $R$, $S$, $Z$, $p(S)$ and $\psi : R \times S \to Z$. Notice that in that definition the only variable not assumed given is the conditional distribution $p(R \mid S)$ which specifies the regulator's responsiveness to the system. The reason for this is that, in our present context, the whole point of regulator design is the specification of this conditional probability distribution. That is, we are assuming that we, as regulator designers, have total control over choosing this distribution in our search for an optimal regulator and also that this is really the only thing we can control. In other words, we are assuming that someone has plunked down onto our workbench some system with a behavioral repertoire $S$ and an associated probability distribution $p(S)$, along with some regulator with a behavioral repertoire $R$, a set of possible results $Z$, and a mapping $\psi : R \times S \to Z$, and that we have been asked to find a way to make $R$ responsive to $S$ – that is, to design a conditional distribution $p(R \mid S)$ – in such a way as to minimize the associated value of the entropy function $H(Z)$. Of course, this raises the question as to how we are supposed to accomplish this. We will return to this question shortly.

Actually, C&A make an additional implicit assumption which I want to go ahead and make explicit. It's a small but important detail, and I want to get it out of the way. This assumption is that the system's behavioral repertoire only contains behaviors that the system actually might perform. Another way to say this is that $S$ is such that every probability in $p(S)$ is greater than zero. This is easy enough to do. If we start off with some version of $S$ that contains a behavior, say $s_\alpha$, that cannot occur, that is, which is such that $p(s_\alpha) = 0$, then we can just create a new version of $S$ from which $s_\alpha$ has been excluded. Since $s_\alpha$ could never occur anyway, its exclusion from $S$ has no impact beyond making every probability in $p(S)$ greater than zero and this is necessary if we want to define $p(R|S)$ without having to resort to any special notational gymnastics, and that's why I want to make this assumption explicit.

## A "Useful and Fundamental Property of the Entropy Function"

Generally speaking, the entropy function has a number of properties that make it a useful measure of unpredictability. First of all, for an arbitrary process with event repertoire $X = \{x_1, x_2, \ldots, x_{|X|}\}$ the entropy function $H(X)$ can take on values that range between an absolute minimum of $H(X) = 0$ and an absolute maximum of $H(X) = \log|X|$. When $H(X) = 0$ then the process is said to be completely predictable and when $H(X) = \log|X|$ the process is said to be completely unpredictable. A situation in which $0 < H(X) < \log|X|$ is somewhere between these two extremes.

The entropy function has another attribute that we will need shortly. The authors describe this characteristic of the entropy function as a "useful and fundamental property" which is that if we pick any two probabilities used to calculate the entropy, and we then increase the imbalance between those two probabilities, that is, if we make the bigger one even bigger and the smaller one even smaller, and then if we recalculate the entropy using these new probabilities, then the recalculated entropy will be smaller than the original entropy.

I am going to use a little basic calculus now to prove that this property holds, but if you aren't comfortable with calculus you can just skip this next part and take it on faith that this truly is a property of the entropy function. If you are comfortable with calculus, the argument proceeds as follows.

First of all we start with any set of numbers, say $p_1, p_2, ..., p_n$, where each is assumed to be between 0 and 1 inclusive, and then we choose any two of them, say $p_a$ and $p_b$, where we can assume, without loss of generality, that $p_a \geq p_b$ (otherwise we just re-label them). Then we choose any positive number $\delta$, where $0 < \delta < 1$, such that $1 \geq p_a + \delta > p_a \geq p_b > p_b - \delta \geq 0$ (which actually implies that $0 < \delta < p_b$). Then the claim we wish to prove is the following:

$$H(p_1, p_2, ..., p_a + \delta, ..., p_b - \delta, ...p_n) < H(p_1, p_2, ..., p_a, ..., p_b, ...p_n)$$

In order to prove this result, we hold constant the $p_1, p_2, ..., p_n$ and define the following function $H(\delta)$, where $0 \leq \delta < p_b$,

$$H(\delta) = H(p_1, p_2, ..., p_a + \delta, ..., p_b - \delta, ...p_n)$$

$$= -(p_a + \delta)\log(p_a + \delta) - (p_b - \delta)\log(p_b - \delta) - \sum_{\substack{i=1, \\ i \neq a, \\ i \neq b}}^{n} p_i \log p_i$$

Notice that when $\delta = 0$, this function $H(\delta)$ is equal to the original entropy of

$$H(0) = H(p_1, p_2, ..., p_a + 0, ..., p_b - 0, ...p_n) = H(p_1, p_2, ..., p_a, ..., p_b, ...p_n)$$

Next, still holding constant the $p_1, p_2, ..., p_n$, we take the derivative of $H(\delta)$ with respect to $\delta$:

$$\frac{dH}{d\delta} = \frac{-(p_a + \delta)}{(p_a + \delta)} - \log(p_a + \delta) + \frac{-(p_b - \delta)(-1)}{(p_b - \delta)} - (-1)\log(p_b - \delta)$$

$$= -1 - \log(p_a + \delta) + 1 + \log(p_b - \delta)$$

$$= \log\frac{(p_b - \delta)}{(p_a + \delta)}$$

Now we observe that $\frac{dH}{d\delta} < 0$ for any $\delta$ such that $0 < \delta < p_b$. How do we know this? Well, first of all, remember that we assumed that $p_a \geq p_b$ and so $0 < \delta < p_b$

implies that $p_a + \delta > p_b - \delta$ and thus that $\frac{p_b - \delta}{p_a + \delta} < 1$ and since the logarithm function is strictly increasing over its entire domain we can conclude that $\frac{dH}{d\delta} = \log\left(\frac{p_b - \delta}{p_a + \delta}\right) < \log 1 = 0$. The fact that $\frac{dH}{d\delta} < 0$ when $0 < \delta < p_b$ means that the function *H* is strictly decreasing on the interval $0 < \delta < p_b$ which means that

$$\begin{aligned} H(\delta) &= H(p_1, p_2, ..., p_a + \delta, ..., p_b - \delta, ...p_n) \\ &< H(p_1, p_2, ..., p_a, ..., p_b, ...p_n) \\ &= H(p_1, p_2, ..., p_a + 0, ..., p_b - 0, ...p_n) \\ &= H(0) \end{aligned}$$

.

The important thing to notice in all of that is the following fact:

$$H(p_1, p_2, ..., p_a + \delta, ..., p_b - \delta, ...p_n) < H(p_1, p_2, ..., p_a, ..., p_b, ...p_n)$$

Assuming always, of course, that $0 < \delta < p_b$ and that $p_a \geq p_b$, as stated earlier, meaning, as claimed, that if we *increase* the imbalance between any two probabilities, we cause a *decrease* in the entropy.

## Expressing *H(Z)* in terms of *P(R|S)*

Let's get back to the question we asked a few pages ago. That is, how are we supposed to find a conditional distribution $p(R|S)$ that will minimize $H(Z)$, given the various above mentioned assumptions? Well, the first thing we need in order to answer this question is some way of relating the entropy function to the conditional distribution $p(R|S)$. We will now derive this relationship, but in order to guide our derivation and to make it a bit more concrete we will use the particular example we have already been using. That is, we will use the following for $R$, $S$, $Z$, $p(S)$ and $\psi : R \times S \to Z$:

$$R = \{r_1, r_2, r_3, r_4, r_5\}$$

$$S = \{s_1, s_2, s_3, s_4, s_5\}$$

$$Z = \{ z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8 \}$$

$$p(S) = \{ p(s_1), p(s_2), p(s_3), p(s_4), p(s_5) \}$$

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ |
| $r_2$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ | $z_2$ |
| $r_3$ | $z_3$ | $z_7$ | $z_5$ | $z_6$ | $z_7$ |
| $r_4$ | $z_8$ | $z_2$ | $z_2$ | $z_1$ | $z_4$ |
| $r_5$ | $z_3$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ |

Now, as mentioned earlier on, once any regulator has been built to exhibit the various behaviors in $R$, be it an optimal regulator or otherwise, the act of setting it up to respond to that system corresponds to the specification of the conditional probability distribution $p(R|S)$ which we can reproduce in its general form as it applies to our example as follows:

| $p(R\|S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $p(r_1\|s_1)$ | $p(r_1\|s_2)$ | $p(r_1\|s_3)$ | $p(r_1\|s_4)$ | $p(r_1\|s_5)$ |
| $r_2$ | $p(r_2\|s_1)$ | $p(r_2\|s_2)$ | $p(r_2\|s_3)$ | $p(r_2\|s_4)$ | $p(r_2\|s_5)$ |
| $r_3$ | $p(r_3\|s_1)$ | $p(r_3\|s_2)$ | $p(r_3\|s_3)$ | $p(r_3\|s_4)$ | $p(r_3\|s_5)$ |
| $r_4$ | $p(r_4\|s_1)$ | $p(r_4\|s_2)$ | $p(r_4\|s_3)$ | $p(r_4\|s_4)$ | $p(r_4\|s_5)$ |
| $r_5$ | $p(r_5\|s_1)$ | $p(r_5\|s_2)$ | $p(r_5\|s_3)$ | $p(r_5\|s_4)$ | $p(r_5\|s_5)$ |

Remember that each entry $p(r_i|s_j)$ in the table represents a number that is understood to be the conditional probability that the regulator responds to an occurrence of $s_j \in S$ with the response $r_i \in R$. Remember also that because these are probabilities, each must be a number between zero and one, and because they are conditional probabilities (each being conditioned on a particular column) we require that the sum of the numbers in any given column must equal exactly one. More formally, we can write:

$$0 \le p(r_i|s_j) \le 1, \text{ for each } i = 1,2,\ldots,|R| \text{ , and } j = 1,2,\ldots,|S|; \text{ and}$$

$$\sum_{i=1}^{|R|} p(r_i|s_j) = p(r_1|s_j) + p(r_2|s_j) + \cdots + p(r_{|R|}|s_j) = 1, \text{ for each } j = 1,2,\ldots,|S|.$$

Of course, for our particular example we have that $|S| = 5$ and $|R| = 5$.

Now, as it concerns our particular example, one of the infinite ways we might assign actual numbers to this distribution is as we did earlier:

| $p(R \mid S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | .01 | .05 | 0 | 0 | .25 |
| $r_2$ | .08 | .09 | 0 | .50 | .39 |
| $r_3$ | .60 | .15 | 0 | .50 | .06 |
| $r_4$ | .25 | .40 | 1 | 0 | 0 |
| $r_5$ | .06 | .31 | 0 | 0 | .30 |

One important property to notice about such conditional probability distributions that we will make use of a little later on is that as long as we respect the two basic conditions (that each number in the table is between zero and one and that the numbers in any column sum to one) then we are free to take any valid such distribution $p(R \mid S)$ and play around with any one of its columns in any way we like to create a different distribution, say $p'(R \mid S)$, that is equally valid. Using our example above, we can, for example, change its third column as shown here:

| $p'(R \mid S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | .01 | .05 | .26 | 0 | .25 |
| $r_2$ | .08 | .09 | .2 | .50 | .39 |
| $r_3$ | .60 | .15 | .15 | .50 | .06 |
| $r_4$ | .25 | .40 | .15 | 0 | 0 |
| $r_5$ | .06 | .31 | .42 | 0 | .30 |

And the resulting distribution, which I've called $p'(R \mid S)$, is equally valid, meaning that it still respects the two basic conditions. The point here is that whenever we change a column in such a table, we only have to worry about the numbers in that particular column. Another way to say this is that each column is independent of the other columns.

Now, with the distribution $p(S)$ assumed given, the specification of the conditional distribution $p(R \mid S)$ determines the *joint* distribution $p(R,S)$, since for any $s_j \in S$ and $r_i \in R$ we have that the probability that both of these occur together is $p(r_i, s_j) = p(s_j)p(r_i \mid s_j)$ for each $i = 1, 2, \ldots, |R|$, and $j = 1, 2, \ldots, |S|$. For our particular example, we can represent this joint probability distribution for $R$ and $S$ in its most general form as follows:

| $p(R,S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $p(r_1,s_1)$ | $p(r_1,s_2)$ | $p(r_1,s_3)$ | $p(r_1,s_4)$ | $p(r_1,s_5)$ |
| $r_2$ | $p(r_2,s_1)$ | $p(r_2,s_2)$ | $p(r_2,s_3)$ | $p(r_2,s_4)$ | $p(r_2,s_5)$ |
| $r_3$ | $p(r_3,s_1)$ | $p(r_3,s_2)$ | $p(r_3,s_3)$ | $p(r_3,s_4)$ | $p(r_3,s_5)$ |
| $r_4$ | $p(r_4,s_1)$ | $p(r_4,s_2)$ | $p(r_4,s_3)$ | $p(r_4,s_4)$ | $p(r_4,s_5)$ |
| $r_5$ | $p(r_5,s_1)$ | $p(r_5,s_2)$ | $p(r_5,s_3)$ | $p(r_5,s_4)$ | $p(r_5,s_5)$ |

But the specification of $p(R|S)$ *also* determines the probability distribution $p(Z) = \left\{ p(z_1), p(z_2), \ldots, p\left(z_{|Z|}\right) \right\}$ and thus the entropy function for $H(Z)$.

How so? To see this, let's consider our example. Take another look at the table for our example mapping $\psi : R \times S \to Z$, which I'll reproduce here:

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ |
| $r_2$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ | $z_2$ |
| $r_3$ | $z_3$ | $z_7$ | $z_5$ | $z_6$ | $z_7$ |
| $r_4$ | $z_8$ | $z_2$ | $z_2$ | $z_1$ | $z_4$ |
| $r_5$ | $z_3$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ |

Notice that in this table, the element $z_6$ appears three times; once each in the columns $s_1$, $s_2$ and $s_4$. That is, $\psi(r_2,s_1) = \psi(r_3,s_4) = \psi(r_5,s_2) = z_6$. Given this information along with the probabilities in the joint distribution $p(R,S)$, we can calculate $p(z_6)$, the probability of the outcome $z_6$. In order to calculate $p(z_6)$ we first recognize that each of the $s_j \in S$ are mutually exclusive as are each of the $r_i \in R$ which means that each of the events $\{r_2,s_1\}$, $\{r_3,s_4\}$ and $\{r_5,s_2\}$ are also mutually exclusive. Furthermore, the event $\{z_6\}$ occurs if and only if any one of the mutually exclusive events $\{r_2,s_1\}$, $\{r_3,s_4\}$ or $\{r_5,s_2\}$ occurs. And this means that we can calculate $p(z_6)$ as the simple sum of each of the probabilities $p(r_2,s_1)$, $p(r_3,s_4)$ and $p(r_5,s_2)$. That is,

$$p(z_6) = p(r_2,s_1) + p(r_3,s_4) + p(r_5,s_2)$$

Reasoning in the same way for each of the elements of $Z$ yields the following for the distribution $p(Z)$:

$$p(Z) = \left\{ \begin{array}{l} p(z_1) = p(r_1,s_1) + p(r_2,s_4) + p(r_4,s_4) + p(r_5,s_5), \\ p(z_2) = p(r_1,s_2) + p(r_2,s_5) + p(r_4,s_2) + p(r_4,s_3), \\ p(z_3) = p(r_1,s_3) + p(r_3,s_1) + p(r_5,s_1), \\ p(z_4) = p(r_1,s_4) + p(r_4,s_5), \\ p(z_5) = p(r_1,s_5) + p(r_3,s_3), \\ p(z_6) = p(r_2,s_1) + p(r_3,s_4) + p(r_5,s_2), \\ p(z_7) = p(r_2,s_2) + p(r_3,s_2) + p(r_3,s_5) + p(r_5,s_3), \\ p(z_8) = p(r_2,s_3) + p(r_4,s_1) + p(r_5,s_4) \end{array} \right\}$$

Now, equipped with this probability distribution for the elements of *Z*, we can now calculate the entropy of $Z$ for our particular example as follows:

$$\begin{aligned} H(Z) &= -\sum_{k=1}^{8} p(z_k)\log p(z_k) = -p(z_1)\log p(z_1) - \cdots - p(z_8)\log p(z_8) \\ &= -\left[p(r_1,s_1) + p(r_2,s_4) + p(r_4,s_4) + p(r_5,s_5)\right]\log\left[p(r_1,s_1) + p(r_2,s_4) + p(r_4,s_4) + p(r_5,s_5)\right] \\ &\quad -\left[p(r_1,s_2) + p(r_2,s_5) + p(r_4,s_2) + p(r_4,s_3)\right]\log\left[p(r_1,s_2) + p(r_2,s_5) + p(r_4,s_2) + p(r_4,s_3)\right] \\ &\quad -\left[p(r_1,s_3) + p(r_3,s_1) + p(r_5,s_1)\right]\log\left[p(r_1,s_3) + p(r_3,s_1) + p(r_5,s_1)\right] \\ &\quad -\left[p(r_1,s_4) + p(r_4,s_5)\right]\log\left[p(r_1,s_4) + p(r_4,s_5)\right] \\ &\quad -\left[p(r_1,s_5) + p(r_3,s_3)\right]\log\left[p(r_1,s_5) + p(r_3,s_3)\right] \\ &\quad -\left[p(r_2,s_1) + p(r_3,s_4) + p(r_5,s_2)\right]\log\left[p(r_2,s_1) + p(r_3,s_4) + p(r_5,s_2)\right] \\ &\quad -\left[p(r_2,s_2) + p(r_3,s_2) + p(r_3,s_5) + p(r_5,s_3)\right]\log\left[p(r_2,s_2) + p(r_3,s_2) + p(r_3,s_5) + p(r_5,s_3)\right] \\ &\quad -\left[p(r_2,s_3) + p(r_4,s_1) + p(r_5,s_4)\right]\log\left[p(r_2,s_3) + p(r_4,s_1) + p(r_5,s_4)\right] \end{aligned}$$

It will be useful later to be able to write these equations out for the general case. The problem is that we cannot predict which of the $p(r_i,s_j)$ will find their way into

the summation for a given $p(z_k)$ and so we cannot use the standard index method to represent the summation. That is, in the general case, we *cannot* write anything like the following:

$$p(z_k) = \sum_{j=1}^{M} \sum_{i=1}^{N} p(r_i, s_j)$$

However, one way around this is as follows:

$$p(z_k) = \sum_{\psi^{-1}(z_k)} p(r,s)$$

Where $\psi^{-1}(z_k)$ is the set of all and only those ordered pairs $\langle r,s \rangle \in R \times S$ that map to $z_k \in Z$ under $\psi$. More formally, $\psi^{-1}(z_k) \equiv \{\langle r,s \rangle \in R \times S : \psi(r,s) = z_k\}$.

Furthermore, if want to indicate that we are only summing the joint probabilities for, say, the particular column $s_j$, we can use the following notational device:

$$p(z_k) = \sum_{\psi_{s_j}^{-1}(z_k)} p(r,s_j)$$

Where $\psi_{s_j}^{-1}(z_k)$ is the set of all and only those ordered pairs $\langle r,s_j \rangle \in R \times S$ that map to $z_k \in Z$ under $\psi$, for some *particular* $s_j \in S$. More formally, $\psi_{s_j}^{-1}(z_k) \equiv \{\langle r,s_j \rangle \in R \times S : \psi(r,s_j) = z_k\}$. Note that for any two distinct $s_j, s_h \in S$, the sets $\psi_{s_j}^{-1}(z_k)$ and $\psi_{s_h}^{-1}(z_k)$ are mutually exclusive. Also, the union of all such sets over $S$ is just the set $\psi^{-1}(z_k)$. Putting this formally, we have

$$\psi^{-1}(z_k) = \psi_{s_1}^{-1}(z_k) \cup \psi_{s_2}^{-1}(z_k) \cup \cdots \cup \psi_{s_{|S|}}^{-1}(z_k) = \bigcup_{j=1}^{|S|} \psi_{s_j}^{-1}(z_k)$$

Using this latter notation we can now write $p(z_k)$ as a double sum using a hybrid indexing style as follows:

$$p(z_k) = \sum_{j=1}^{|S|} \sum_{\psi_{s_j}^{-1}(z_k)} p(r, s_j)$$

This notation allows us to write $p(z_k)$ in terms of the conditional distribution $p(R \mid S)$ as follows:

$$p(z_k) = \sum_{j=1}^{|S|} \sum_{\psi_{s_j}^{-1}(z_k)} p(r, s_j) = \sum_{j=1}^{|S|} \sum_{\psi_{s_j}^{-1}(z_k)} p(r, s_j) \frac{p(s_j)}{p(s_j)}$$
$$= \sum_{j=1}^{|S|} p(s_j) \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j)$$

Let's take another look at that last expression. In particular, let's focus on the expression used in the inner sum:

$$\sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j)$$

The first thing to realize about this expression is that it is the *conditional* probability that the regulator produces the outcome $z_k$ in response to $s_j \in S$. That is,

$$\sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j) = p(z_k \mid s_j)$$

Secondly, in the general case, and for every $s_j \in S$, this quantity is always between zero and one (inclusively). That is,

$$0 \le \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j) = p(z_k \mid s_j) \le 1$$

Thirdly, for any particular $s_j \in S$, it may be the case either that there are no occurrences of $z_k$ in the $s_j$ column in which case $\psi_{s_j}^{-1}(z_k) = \emptyset$ or that for every occurrence of $z_k$ in the $s_j$ column $p(r \mid s_j) = 0$. If either of these is true then we would have, for that particular $s_j$,

$$\sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j) = p(z_k \mid s_j) = 0$$

And fourth, again for any particular $s_j \in S$, it may be the case that $z_k$ is the *only* outcome in the $s_j$ column for which $p(r \mid s_j) > 0$. When this is true then we would have, for that particular $s_j$,

$$\sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j) = p(z_k \mid s_j) = 1$$

To illustrate all of this with our particular example we would have the following sets for $\psi^{-1}(z_k)$, for $k = 1,2,\ldots,|Z|$.

$$\begin{aligned}
\psi^{-1}(z_1) &= \{\langle r_1, s_1\rangle, \langle r_2, s_4\rangle, \langle r_4, s_4\rangle, \langle r_5, s_5\rangle\} \\
\psi^{-1}(z_2) &= \{\langle r_1, s_2\rangle, \langle r_2, s_5\rangle, \langle r_4, s_2\rangle, \langle r_4, s_3\rangle\} \\
\psi^{-1}(z_3) &= \{\langle r_1, s_3\rangle, \langle r_3, s_1\rangle, \langle r_5, s_1\rangle\} \\
\psi^{-1}(z_4) &= \{\langle r_1, s_4\rangle, \langle r_4, s_5\rangle\} \\
\psi^{-1}(z_5) &= \{\langle r_1, s_5\rangle, \langle r_3, s_3\rangle\} \\
\psi^{-1}(z_6) &= \{\langle r_2, s_1\rangle, \langle r_3, s_4\rangle, \langle r_5, s_2\rangle\} \\
\psi^{-1}(z_7) &= \{\langle r_2, s_2\rangle, \langle r_3, s_2\rangle, \langle r_3, s_5\rangle, \langle r_5, s_3\rangle\} \\
\psi^{-1}(z_8) &= \{\langle r_2, s_3\rangle, \langle r_4, s_1\rangle, \langle r_5, s_4\rangle\}
\end{aligned}$$

And as for the $\psi_{s_j}^{-1}(z_k)$, for $k = 1,2,\ldots,|Z|$ we have:

$$\psi_{s_1}^{-1}(z_1) = \{\langle r_1, s_1\rangle\} \qquad \psi_{s_1}^{-1}(z_2) = \{\ \} \qquad \psi_{s_1}^{-1}(z_3) = \{\langle r_3, s_1\rangle, \langle r_5, s_1\rangle\}$$
$$\psi_{s_2}^{-1}(z_1) = \{\ \} \qquad \psi_{s_2}^{-1}(z_2) = \{\langle r_1, s_2\rangle, \langle r_4, s_2\rangle\} \qquad \psi_{s_2}^{-1}(z_3) = \{\ \}$$
$$\psi_{s_3}^{-1}(z_1) = \{\ \} \qquad \psi_{s_3}^{-1}(z_2) = \{\langle r_4, s_3\rangle\} \qquad \psi_{s_3}^{-1}(z_3) = \{\langle r_1, s_3\rangle\}$$
$$\psi_{s_4}^{-1}(z_1) = \{\langle r_2, s_4\rangle, \langle r_4, s_4\rangle\} \qquad \psi_{s_4}^{-1}(z_2) = \{\ \} \qquad \psi_{s_4}^{-1}(z_3) = \{\ \}$$
$$\psi_{s_5}^{-1}(z_1) = \{\langle r_5, s_5\rangle\} \qquad \psi_{s_5}^{-1}(z_2) = \{\langle r_2, s_5\rangle\} \qquad \psi_{s_5}^{-1}(z_3) = \{\ \}$$

$$\psi_{s_1}^{-1}(z_4) = \{\ \} \qquad \psi_{s_1}^{-1}(z_5) = \{\ \} \qquad \psi_{s_1}^{-1}(z_6) = \{\langle r_2, s_1\rangle\}$$
$$\psi_{s_2}^{-1}(z_4) = \{\ \} \qquad \psi_{s_2}^{-1}(z_5) = \{\ \} \qquad \psi_{s_2}^{-1}(z_6) = \{\langle r_5, s_2\rangle\}$$
$$\psi_{s_3}^{-1}(z_4) = \{\ \} \qquad \psi_{s_3}^{-1}(z_5) = \{\langle r_3, s_3\rangle\} \qquad \psi_{s_3}^{-1}(z_6) = \{\ \}$$
$$\psi_{s_4}^{-1}(z_4) = \{\langle r_1, s_4\rangle\} \qquad \psi_{s_4}^{-1}(z_5) = \{\ \} \qquad \psi_{s_4}^{-1}(z_6) = \{\langle r_3, s_4\rangle\}$$
$$\psi_{s_5}^{-1}(z_4) = \{\langle r_4, s_5\rangle\} \qquad \psi_{s_5}^{-1}(z_5) = \{\langle r_1, s_5\rangle\} \qquad \psi_{s_5}^{-1}(z_6) = \{\ \}$$

$$\psi_{s_1}^{-1}(z_7) = \{\ \} \qquad \psi_{s_1}^{-1}(z_8) = \{\langle r_4, s_1\rangle\}$$
$$\psi_{s_2}^{-1}(z_7) = \{\langle r_2, s_2\rangle, \langle r_3, s_2\rangle\} \qquad \psi_{s_2}^{-1}(z_8) = \{\ \}$$
$$\psi_{s_3}^{-1}(z_7) = \{\langle r_5, s_3\rangle\} \qquad \psi_{s_3}^{-1}(z_8) = \{\langle r_2, s_3\rangle\}$$
$$\psi_{s_4}^{-1}(z_7) = \{\ \} \qquad \psi_{s_4}^{-1}(z_8) = \{\langle r_5, s_4\rangle\}$$
$$\psi_{s_5}^{-1}(z_7) = \{\langle r_3, s_5\rangle\} \qquad \psi_{s_5}^{-1}(z_8) = \{\ \}$$

Equipped with these more compact notational devices we can write the general form of the entropy function in terms of both the joint distribution $p(R,S)$ and the conditional distribution $p(R \mid S)$ as follows:

$$H(Z) = -\sum_{k=1}^{|Z|} p(z_k) \log p(z_k) = -\sum_{k=1}^{|Z|} \left[ \overbrace{\sum_{\psi^{-1}(z_k)} p(r,s)}^{\text{this is } p(z_k)} \right] \log \left[ \overbrace{\sum_{\psi^{-1}(z_k)} p(r,s)}^{\text{this is } p(z_k)} \right]$$

$$= -\sum_{k=1}^{|Z|} \left[ \overbrace{\sum_{j=1}^{|S|} p(s_j) \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j)}^{\text{this is } p(z_k)} \right] \log \left[ \overbrace{\sum_{j=1}^{|S|} p(s_j) \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j)}^{\text{this is } p(z_k)} \right]$$

Note that this is the *general* form of the entropy function in terms of $p(R,S)$ and $p(R \mid S)$ not simply the version that pertains to our example. (To obtain the version that applies to our particular example we would have only to plug in the values $|Z| = 8$, $|R| = 5$ and $|S| = 5$). And equipped with this expression for the entropy function we are reading to return to the question of how to go about designing $p(R \mid S)$ so that $H(Z)$ is made as small as possible (given $R$, $S$, $Z$, $p(S)$ and $\psi : R \times S \to Z$).

## A Lemma Regarding Successful Regulators

The next thing we need to accomplish this is to define, as do C&A, the set $\pi$ that contains exactly the sort of conditional probability distributions $p(R \mid S)$ that we are looking for – that is, that make the resulting entropy function $H(Z)$ as small as possible. Actually, $\pi$ is not really as simple as that because another point the authors make is that it is possible for two different $p(R \mid S)$ distributions, say $p_1(R \mid S)$ and $p_2(R \mid S)$ to determine two different $p(Z)$ distributions, say $p_1(Z)$ and $p_2(Z)$, where $p_1(Z) \neq p_2(Z)$, in such a way that the resulting entropies $H_1(Z)$ and $H_2(Z)$ are equal. They claim that to consider this possibility would complicate unnecessarily their proof and so they place the additional restriction on $\pi$ that each of its members determines (via the given $p(S)$ and $\psi$) the same unique $p(Z)$. To make this point more vivid, let's call the minimum attainable entropy $H_{\min}(Z)$. Now, if we discover that, say, both $p_1(R \mid S)$ and $p_2(R \mid S)$ achieve $H_{\min}(Z)$, but that $p_1(R \mid S)$ determines $p_1(Z)$ and that $p_2(R \mid S)$ determines $p_2(Z)$ where $p_1(Z) \neq p_2(Z)$, then what the authors want to do is to exclude from the set $\pi$ either $p_1(R \mid S)$ or $p_2(R \mid S)$

Thus, $\pi$ does not contain *all* of the distributions $p(R|S)$ that achieve $H_{\min}(Z)$. The specification of $\pi$ requires that we first pick one of the $p(Z)$ distributions that achieves $H_{\min}(Z)$, and that we include only those $p(R|S)$ distributions in $\pi$ that achieve the chosen $p(Z)$ distribution that in turn determines $H_{\min}(Z)$. I would like to examine this simplification in more detail, but not at this point. For now, let's assume the trick is valid and choose an arbitrary $p(R|S)$ from $\pi$.

Now we come to a result that Conant and Ashby call a lemma and which they describe as the heart of their proof. To put it simply, this lemma tells us that the only way to build an optimal regulator (one that makes $H(Z)$ minimal) is to make sure that it never acts so as to produces different outcomes on different occasions when confronted with the same system behavior. The way they prove this claim is by the method of contradiction, which means that they start by assuming the contrary claim – that it is possible to build an optimal regulator that might produce different outcomes when confronted on different occasions with the same system behavior – and then they deduce a logical contradiction. The deduction of the logical contradiction under the assumption that the original claim is false proves that the original claim is true. We will now walk through Conant and Ashby's proof of their lemma.

Perhaps the most difficult thing to understand about this lemma is the symbol-dense mathematically terse way they state what it claims which resembles the following:

*Lemma* : For an arbitrary $p(R|S)$ in $\pi$ and for all $s_j \in S$, the set $\{\psi(r_i, s_j) : p(r_i, s_j) > 0\}$ has only one element.

I think the best way to unpack what this lemma is saying is as I already did in the previous paragraph, but in case the connection isn't obvious we can start with the following. What the above lemma is claiming, is that provided we are looking at an optimal conditional distribution $p(R|S) \in \pi$, then for any given system behavior $s_j \in S$, the conditional distribution $p(R|S = s_j)$ is such that if $p(r_h|s_j) > 0$ and $p(r_i|s_j) > 0$ for any $r_h, r_i \in R$, $r_h \neq r_i$, then it must be the case that $\psi(r_h, s_j) = \psi(r_i, s_j) = z_k$, for some $z_k \in Z$. In case that is still too abstruse, let's try walking through it using our specific example. Remember that our goal is to specify the following table:

| $p(R \mid S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $p(r_1 \mid s_1)$ | $p(r_1 \mid s_2)$ | $p(r_1 \mid s_3)$ | $p(r_1 \mid s_4)$ | $p(r_1 \mid s_5)$ |
| $r_2$ | $p(r_2 \mid s_1)$ | $p(r_2 \mid s_2)$ | $p(r_2 \mid s_3)$ | $p(r_2 \mid s_4)$ | $p(r_2 \mid s_5)$ |
| $r_3$ | $p(r_3 \mid s_1)$ | $p(r_3 \mid s_2)$ | $p(r_3 \mid s_3)$ | $p(r_3 \mid s_4)$ | $p(r_3 \mid s_5)$ |
| $r_4$ | $p(r_4 \mid s_1)$ | $p(r_4 \mid s_2)$ | $p(r_4 \mid s_3)$ | $p(r_4 \mid s_4)$ | $p(r_4 \mid s_5)$ |
| $r_5$ | $p(r_5 \mid s_1)$ | $p(r_5 \mid s_2)$ | $p(r_5 \mid s_3)$ | $p(r_5 \mid s_4)$ | $p(r_5 \mid s_5)$ |

And that we want to specify this table so that the resulting entropy $H(Z)$ is at a minimum. Now, suppose we have done just that, so that the above table contains a bunch of numbers, all of which are between 0 and 1 such that the sum of the numbers in any given column equals 1 and that if we use these number to calculate $H(Z)$ then the number we come up with will be the smallest we could achieve. Now, the lemma is telling us that if we first focus our attention on any of the $s_j \in S$, which we can represent as follows:

| $p(R \mid S)$ | $\cdots$ | $s_j$ | $\cdots$ |
|---|---|---|---|
| $r_1$ | $\cdots$ | $p(r_1 \mid s_j)$ | $\cdots$ |
| $r_2$ | $\cdots$ | $p(r_2 \mid s_j)$ | $\cdots$ |
| $r_3$ | $\cdots$ | $p(r_3 \mid s_j)$ | $\cdots$ |
| $r_4$ | $\cdots$ | $p(r_4 \mid s_j)$ | $\cdots$ |
| $r_5$ | $\cdots$ | $p(r_5 \mid s_j)$ | $\cdots$ |

And if we then notice that $p(r_h \mid s_j) > 0$ and $p(r_i \mid s_j) > 0$ which we can visualize as follows:

| $p(R \mid S)$ | $\cdots$ | $s_j$ | $\cdots$ |
|---|---|---|---|
| $\vdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $r_h$ | $\cdots$ | $p(r_h \mid s_j) > 0$ | $\cdots$ |
| $\vdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $r_i$ | $\cdots$ | $p(r_i \mid s_j) > 0$ | $\cdots$ |
| $\vdots$ | $\cdots$ | $\cdots$ | $\cdots$ |

Then according to the lemma, it must be the case that $\psi\left(r_h, s_j\right) = \psi\left(r_i, s_j\right) = z_k$ for some particular $z_k \in Z$. That is, our table for $\psi$ must look as follows:

| $\psi$ | $\cdots$ | $s_j$ | $\cdots$ |
|---|---|---|---|
| $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ |
| $r_h$ | $\cdots$ | $z_k$ | $\cdots$ |
| $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ |
| $r_i$ | $\cdots$ | $z_k$ | $\cdots$ |
| $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ |

In other words, and as I explained above, the lemma is telling us that the only way to build an optimal regulator (one that makes $H(Z)$ minimal) is to make sure that it *never* produces two different outcomes when confronted with the same $s_j \in S$. That is, given some particular $s_j \in S$, when faced with any two responses, say $r_x$ and $r_y$, that might otherwise produce different outcomes, then at least one of them should have a probability of zero; that is, if $\psi\left(r_x, s_j\right) \neq \psi\left(r_y, s_j\right)$ then we better make sure that either $p(r_x \mid s_j) = 0$ or $p(r_y \mid s_j) = 0$, because otherwise $H(Z)$ cannot be minimal.

In any case, that's what the lemma is claiming. And as I said above, the authors use a proof by contradiction to establish this result. That is, they first assume, contrary to what the lemma claims is possible, that there does exist a conditional distribution $p(R \mid S) \in \pi$ (i.e. that minimizes $H(Z)$) which is such that there exists some $s_j \in S$ such that $p\left(r_x \mid s_j\right) > 0$ and $p\left(r_y \mid s_j\right) > 0$, for $r_x, r_y \in R$, $r_x \neq r_y$, where the outcomes under $\psi$ are *different*, that is, $\psi\left(r_x, s_j\right) = z_k \neq z_l = \psi\left(r_y, s_j\right)$, for some $z_k, z_l \in Z$. After making this assumption, they then deduce a contradiction, which implies that their assumption could not have been true, which, in turn, implies that the lemma must be true.

The details of the proof are as follows. Let's review the assumptions. First we are assuming, contrary to what the lemma claims is possible, that there does exist a conditional distribution $p(R \mid S) \in \pi$ (i.e. that minimizes $H(Z)$) which is such that there exists some $s_j \in S$ such that $p\left(r_x \mid s_j\right) > 0$ and $p\left(r_y \mid s_j\right) > 0$, for $r_x, r_y \in R$, $r_x \neq r_y$, where the outcomes under $\psi$ are *different*, that is, $\psi\left(r_x, s_j\right) = z_k \neq z_l = \psi\left(r_y, s_j\right)$, for some $z_k, z_l \in Z$.

Now, let's consider $p(z_k)$ and $p(z_l)$. Using the notation we worked out earlier we can write each of these as simple sums of the probabilities in the distribution of $p(R,S)$. That is, we can write

$$p(z_k) = \sum_{\psi^{-1}(z_k)} p(r,s)$$

and

$$p(z_l) = \sum_{\psi^{-1}(z_l)} p(r,s)$$

But we are assuming that $\psi(r_x,s_j) = z_k$ and $\psi(r_y,s_j) = z_l$ which means that

$$\langle r_x,s_j\rangle \in \psi^{-1}(z_k)$$

and

$$\langle r_y,s_j\rangle \in \psi^{-1}(z_l)$$

Therefore we can write

$$p(z_k) = p(r_x,s_j) + \sum_{\psi^{-1}(z_k)-\langle r_x,s_j\rangle} p(r,s)$$

and

$$p(z_l) = p(r_y,s_j) + \sum_{\psi^{-1}(z_l)-\langle r_y,s_j\rangle} p(r,s)$$

Where we are using the notation $\psi^{-1}(z_k) - \langle r_x,s_j\rangle$ to represent the set of all ordered pairs $\langle r,s\rangle \in R \times S$ such that $\psi(r,s) = z_k$ but that *excludes* the particular ordered pair $\langle r_x,s_j\rangle$. Similarly the notation $\psi^{-1}(z_l) - \langle r_y,s_j\rangle$ represents the set of all ordered pairs $\langle r,s\rangle \in R \times S$ such that $\psi(r,s) = z_l$ but that *excludes* the particular ordered pair $\langle r_y,s_j\rangle$. More formally we can write,

$$\psi^{-1}(z_k) - \langle r_x, s_j \rangle = \{\langle r,s \rangle \in R \times S : \psi(r,s) = z_k, \langle r,s \rangle \neq \langle r_x, s_j \rangle\}$$

and

$$\psi^{-1}(z_l) - \langle r_y, s_j \rangle = \{\langle r,s \rangle \in R \times S : \psi(r,s) = z_l, \langle r,s \rangle \neq \langle r_y, s_j \rangle\}$$

.

Because these expressions are a bit bulky, let's make use of the following symbol substitutions:

$$p_k = p(r_x, s_j),$$
$$p_l = p(r_y, s_j),$$
$$\varepsilon_k = \sum_{\psi^{-1}(z_k) - \langle r_x, s_j \rangle} p(r,s),$$

and

$$\varepsilon_l = \sum_{\psi^{-1}(z_l) - \langle r_y, s_j \rangle} p(r,s)$$

Using these substitutions our expressions for $p(z_k)$ and $p(z_l)$ become,

$$p(z_k) = p_k + \varepsilon_k$$

and

$$p(z_l) = p_l + \varepsilon_l$$

With the above streamlined expressions for $p(z_k)$ and $p(z_l)$ we can now express the entropy function as follows:

$$H(Z) = -\overbrace{(p_k + \varepsilon_k)}^{\text{this is } p(z_k)} \log \overbrace{(p_k + \varepsilon_k)}^{\text{this is } p(z_k)} - \overbrace{(p_l + \varepsilon_l)}^{\text{this is } p(z_l)} \log \overbrace{(p_l + \varepsilon_l)}^{\text{this is } p(z_l)} - \sum_{\substack{h=1, \\ h \neq k, \\ h \neq l}}^{|Z|} p(z_h) \log p(z_h) \quad .$$

$$= -\overbrace{(p_k + \varepsilon_k)}^{\text{this is } p(z_k)} \log \overbrace{(p_k + \varepsilon_k)}^{\text{this is } p(z_k)} - \overbrace{(p_l + \varepsilon_l)}^{\text{this is } p(z_l)} \log \overbrace{(p_l + \varepsilon_l)}^{\text{this is } p(z_l)} + \rho$$

Where we have introduced yet another space saving substitution in the above by setting

$$\rho = -\sum_{\substack{h=1,\\ h\neq k,\\ h\neq l}}^{|Z|} p(z_h)\log p(z_h)$$

Now, remember, this is *supposed* to be the smallest possible value for $H(Z)$ that is attainable because we are using $p(R|S)\in\pi$ to calculate it. More specifically, we have used $p(r_x|s_j)$ and $p(r_y|s_j)$ from $p(R|S)$ to calculate both $p(r_x,s_j)$ and $p(r_y,s_j)$ (since $p(r_x,s_j)=p(s_j)\,p(r_x|s_j)$ and $p(r_y,s_j)=p(s_j)\,p(r_y|s_j)$), which we then used to calculate each of $p(z_k)$ and $p(z_l)$. So really, we should write

$$H_{\min}(Z) = -\overbrace{(p_k+\varepsilon_k)}^{\text{this is } p(z_k)}\log\overbrace{(p_k+\varepsilon_k)}^{\text{this is } p(z_k)} - \overbrace{(p_l+\varepsilon_l)}^{\text{this is } p(z_l)}\log\overbrace{(p_l+\varepsilon_l)}^{\text{this is } p(z_l)} + \rho$$

But now if we first figure out which of $p_k$ and $p_l$ is bigger and then we increase the imbalance between them by adding some $\delta$ such that $0<\delta<\min(p_k,p_l)$ to the larger of the two and subtracting the same $\delta$ from the smaller of the two then by the useful and fundamental property of the entropy function mentioned above we conclude that the resulting $H(Z)$ will be yet smaller than $H_{\min}(Z)$, which is, of course, a contradiction! For example, let's just assume that $p_k \geq p_l$. Then by our useful and fundamental property of the entropy function, we conclude that

$$\begin{aligned} H_{\min}(Z) &= -(p_k+\varepsilon_k)\log(p_k+\varepsilon_k) - (p_l+\varepsilon_l)\log(p_l+\varepsilon_l) + \rho \\ &> -(p_k+\delta+\varepsilon_k)\log(p_k+\delta+\varepsilon_k) - (p_l-\delta+\varepsilon_l)\log(p_l-\delta+\varepsilon_l) + \rho \end{aligned}.$$

But this is a contradiction. For the case in which $p_k < p_l$ then we subtract $\delta$ from $p_k$ and add $\delta$ to $p_l$ in order to obtain the same contradiction. Thus, by assuming that the lemma is false, we derive a logical contradiction, which means that the lemma must be true. That is, for any conditional distribution $p(R|S)\in\pi$ (i.e. that minimizes $H(Z)$) and for any $s_j\in S$, if $p(r_x|s_j)>0$ and $p(r_y|s_j)>0$, for $r_x\neq r_y$, then it must be the case that the overall outcomes are the *same*, that is, that

$\psi\left(r_x, s_j\right) = z_k = \psi\left(r_y, s_j\right)$. To summarize this in words, whenever faced with a given $s_j \in S$, any *optimal* regulator must always produce the same outcome. Notice that it might produce this identical outcome with different responses, say $r_x$ and $r_y$, but it can only use such different responses provided they both map with $s_j$ under $\psi$ to the same $z_k \in Z$. Any regulator that responds to some particular $s_j \in S$ in such a way as to produce, say, $z_k \in Z$ on some occasions and, say, $z_l \in Z$ on different occasions will not be able to minimize $H(Z)$.

## The Simplest Optimal Regulator

What we have established so far is that an optimal (i.e. maximally successful) regulator will always produce the same outcome, say $z_k \in Z$, in response to a given $s_j \in S$. The only thing left now is to prove that the *simplest* way to accomplish this is to pick any one of the responses that map with $s_j$ under $\psi$ to that unique outcome $z_k$ and to have the regulator *always* do that response. For example, if it turns out that the responses $r_w, r_x, r_y \in R$ are such that $p\left(r_w \mid s_j\right) > 0$, $p\left(r_x \mid s_j\right) > 0$, $p\left(r_y \mid s_j\right) > 0$ and that $\psi\left(r_w, s_j\right) = \psi\left(r_x, s_j\right) = \psi\left(r_y, s_j\right) = z_k$, and of course that $H(Z)$ is minimized, then what we should do is to pick any one of these responses, say $r_x$, and set $p\left(r_x \mid s_j\right) = 1$, which means, of course, that we have to set $p\left(r_w \mid s_j\right) = p\left(r_y \mid s_j\right) = 0$.

Actually, it's pretty obvious that this would be the simplest thing to do, provided it doesn't increase $H(Z)$. The authors claim that this won't happen, but how can we be sure? One approach is as follows. First let's take an arbitrary $z_k \in Z$, and express $p(z_k)$ using the notation we developed earlier:

$$p(z_k) = \sum_{j=1}^{|S|} p\left(s_j\right) \sum_{\psi_{s_j}^{-1}(z_k)} p\left(r \mid s_j\right)$$

Now, let us focus our attention on an arbitrary $s_h \in S$ and separate its term out from the rest of the summation as follows:

$$p(z_k) = p(s_h) \sum_{\psi_{s_h}^{-1}(z_k)} p(r \mid s_h) + \sum_{\substack{j=1,\\ j \neq h}}^{|S|} p(s_j) \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j)$$

Although this much is true for any $z_k \in Z$ and any $s_h \in S$, let us now consider a more specific case. That is, let us suppose, first of all, that we are looking at an optimal regulator, so that $p(R \mid S) \in \pi$. Furthermore, let us also suppose that $z_k$ is the particular outcome that our regulator always produces when confronted with $s_h$. Under these conditions, it must be the case that

$$\sum_{\psi_{s_h}^{-1}(z_k)} p(r \mid s_h) = 1$$

And thus, that

$$\begin{aligned} p(z_k) &= p(s_h) \sum_{\psi_{s_h}^{-1}(z_k)} p(r \mid s_h) + \sum_{\substack{j=1,\\ j \neq h}}^{|S|} p(s_j) \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j) \\ &= p(s_h)(1) + \sum_{\substack{j=1,\\ j \neq h}}^{|S|} p(s_j) \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j) \\ &= p(s_h) + \sum_{\substack{j=1,\\ j \neq h}}^{|S|} p(s_j) \sum_{\psi_{s_j}^{-1}(z_k)} p(r \mid s_j) \end{aligned}$$

What we have established here is that under the stated conditions $p(z_k)$ is completely independent of the details of the conditional distribution $p(R \mid s_h)$ and so we can re-arrange $p(R \mid s_h)$ however we wish without having any impact on $p(z_k)$. The only conditions we have to respect are that we have to make sure that $0 \leq p(r_i \mid s_h) \leq 1$ for all $r_i \in R$ and also that $\sum_{i=1}^{|R|} p(r_i \mid s_h) = 1$.

Now, the fact that $p(z_k)$ is completely independent of the details of the conditional distribution $p(R \mid s_h)$, along with the fact that the details of $p(R \mid s_h)$ are completely independent of the details of all of the other column distributions $p(R \mid s_j)$, for all

$s_j \in S$, where $s_j \neq s_h$, means that such rearrangements of $p(R \mid s_h)$ have no impact whatsoever on any of the probabilities in the distribution $p(Z)$ and thus that they have no impact whatsoever on $H(Z)$ as the authors claim. Thus we can do as the authors suggest. That is, starting with some optimal distribution $p(R \mid S) \in \pi$, and assuming (as above) that $z_k$ is the outcome produced by this optimal regulator whenever it is confronted with $s_h$, and letting $r_{\alpha_1}, r_{\alpha_2}, \ldots, r_{\alpha_n} \in R$ represent all and only those responses that map with $s_h$ under $\psi$ to $z_k$, we can create a new optimal distribution $p'(R \mid S) \in \pi$ by picking any one of the responses, say $r_{\alpha_1}$, and setting $p(r_{\alpha_1} \mid s_h) = 1$ and also setting $p(r_{\alpha_2} \mid s_h) = p(r_{\alpha_3} \mid s_h) = \cdots = p(r_{\alpha_n} \mid s_h) = 0$. By doing this for each column $s_1, s_2, \ldots, s_{|S|} \in S$ we will have created an optimal distribution that is also maximally simple.

In order to illustrate this with our particular example, let's take yet another look at our table for $\psi$ :

| $\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ |
| $r_2$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ | $z_2$ |
| $r_3$ | $z_3$ | $z_7$ | $z_5$ | $z_6$ | $z_7$ |
| $r_4$ | $z_8$ | $z_2$ | $z_2$ | $z_1$ | $z_4$ |
| $r_5$ | $z_3$ | $z_6$ | $z_7$ | $z_8$ | $z_1$ |

Now, let's suppose we have a merely optimal regulator (not a maximally simple one) that behaves according to the following schedule:

| When confronted with | Regulator produces outcome | By selecting response |
|---|---|---|
| $s_1$ | $z_3$ | $r_3$ 20% of the time or<br>$r_5$ 80% of the time |
| $s_2$ | $z_7$ | $r_2$ 65% of the time or<br>$r_3$ 35% of the time |
| $s_3$ | $z_7$ | $r_5$ 100% of the time |

| When confronted with | Regulator produces outcome | By selecting response |
|---|---|---|
| $s_4$ | $z_1$ | $r_2$ 15% of the time or<br>$r_4$ 85% of the time |
| $s_5$ | $z_7$ | $r_3$ 100% of the time |

Notice that all of the above information can be stored in special hybrid table that combines the tables for $p(R \mid S)$ and $\psi$ as follows:

| $p(R \mid S)\,/\,\psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | 0 | 0 | 0 | 0 | 0 |
| $r_2$ | 0 | $.65/z_7$ | 0 | $.15/z_1$ | 0 |
| $r_3$ | $.20/z_3$ | $.35/z_7$ | 0 | 0 | $1/z_7$ |
| $r_4$ | 0 | 0 | 0 | $.85/z_1$ | 0 |
| $r_5$ | $.80/z_3$ | 0 | $1/z_7$ | 0 | 0 |

Now, if we want to convert this optimal regulator to one that is maximally simple, we just have to fix columns $s_1$, $s_2$ and $s_4$ so that all of the probability is concentrated on a single response where it really doesn't matter which response we pick, provided it is one of the responses that already has a positive probability. Let's just pick response $r_3$ for columns $s_1$ and $s_2$, and response $r_4$ for column $s_4$. Now our hybrid table for the maximally simple optimal regulator looks as follows:

| $p(R \mid S) \,/\, \psi$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | 0 | 0 | 0 | 0 | 0 |
| $r_2$ | 0 | 0 | 0 | 0 | 0 |
| $r_3$ | $1/z_3$ | $1/z_7$ | 0 | 0 | $1/z_7$ |
| $r_4$ | 0 | 0 | 0 | $1/z_1$ | 0 |
| $r_5$ | 0 | 0 | $1/z_7$ | 0 | 0 |

Removing the outcome elements from the above table gives us our pure table for the distribution $p(R \mid S)$ as follows:

| $p(R \mid S)$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $r_1$ | 0 | 0 | 0 | 0 | 0 |
| $r_2$ | 0 | 0 | 0 | 0 | 0 |
| $r_3$ | 1 | 1 | 0 | 0 | 1 |
| $r_4$ | 0 | 0 | 0 | 1 | 0 |
| $r_5$ | 0 | 0 | 1 | 0 | 0 |

And as discussed in the opening of this paper, this sort of distribution $p(R \mid S)$ which consists entirely of ones and zeros specifies a mapping $h : S \to R$. For this current example, the mapping can be represented as follows:

| $h:$ $S$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
|---|---|---|---|---|---|
| $R$ | $r_3$ | $r_3$ | $r_5$ | $r_4$ | $r_3$ |

## *Conclusion: a "Rigorous Theorem"*

Congratulations. If you have made it this far you have successfully understood the proof of Conant and Ashby's Good-Regulator Theorem. In closing I would like to discuss the following assertion that Conant and Ashby make in their paper:

> The first effect of this theorem is to change the status of model-making from optional to compulsory. As we said earlier, model-making has hitherto largely been suggested (for regulating complex dynamic systems) as a possibility: the theorem shows that, in a very wide class (specified in the proof of the theorem), success in regulation implies that a sufficiently similar model must have been built, *whether it was done explicitly, or simply developed as the regulator was improved*. Thus the would be model-maker now has a rigorous theorem to justify his work. [page 97, emphasis added]

This is a strong statement, but it contains some subtlety that needs to be understood. The subtlety revolves around the phrase "whether it was done explicitly, or simply developed as the regulator was improved." To really see what this entails, let's take another look at the statement of the theorem:

*Theorem* : The simplest optimal regulator $R$ of a system $S$ produces behaviors from $R = \left\{ r_1, r_2, \cdots, r_{|R|} \right\}$ which are related to the behaviors in $S = \left\{ s_1, s_2, \cdots s_{|S|} \right\}$ by a mapping $h : S \to R$.

Notice that this statement actually refers to what we can recognize as *two* distinct models. On the one hand there is the actual regulator device $R$ which is a *dynamic* entity and which, through its simplest optimal regulation of the system $S$, comes to *be* a dynamic model of that system. This is the so-called "good-regulator model" referred to by the title of C&A's paper: every good regulator of a system must *be* a model of that system. But the statement of the theorem also refers to the *mapping* $h : S \to R$, which is *also* a model. More specifically, this mapping is a model of the representational relationship between the good-regulator and the system it regulates. This mapping is a model for the same reason that the good-regulator itself is a model, because there is yet *another* mapping, call it $k$, that is set up between the component "bits and pieces" of the regulator/system ensemble and the component "bits and pieces" of the mapping $h : S \to R$. Of course, this mapping we are calling $k$ is just the one we build in our own minds so that we can understand the theorem.

This second model $h : S \to R$ is really just a *description* of the good-regulator model. We can think of it as the "technical specification" of that model. It is a

technical specification in the sense that we could use it to *build* the good-regulator model. It is essentially a look-up table that tells us specifically which of the regulator's behaviors ought to be set up as responses to the system's behaviors so that the regulator device actually regulates the system in the simplest optimal way. We might also call this technical specification a "control-model" because it appears to be controlling the good-regulator model, or at least its construction. But it is important to see that these might only be appearances. The most we can really say is that the regulator acts *as if* it were being guided by the control-model, or perhaps *as if* it were built according to the technical specifications. But the fact that the regulator acts *as if* one or both of these were true does not mean they are, in fact, true. At the extreme, it might simply be the result of some massive coincidence that the good-regulator is behaving this way. As unlikely as this might be, it is a legitimate possibility.

An actual example will make all this clearer. Consider a dedicated father attempting to assemble a ping pong table down in the basement late one Christmas Eve after his children have gone to bed. The system to be regulated here consists of all the various pieces of the table (surface, legs, braces, screws, net, etc.) along with all of the tools needed to put the thing together. The goal, of course, is the assembled table and the good-regulator of this system is the dedicated father. Another key component to all of this is the sheet of assembly instructions that show how to put all of the pieces together into a ping pong table. These assembly instructions correspond to the mapping $h : S \to R$ referred to in the theorem and comprise what we are calling the "technical specifications" or the "control-model". They are "technical specifications" in the sense that the father can use them to "build" himself into a good-regulator model of the system of tools and ping pong table pieces. The instructions are a "control-model" in the sense that the father can use them to guide (control) his own behavior and thus behave as a good-regulator of this system.

Now, the thing to realize here is that C&A's theorem really only proves that the *first* model – i.e. the good-regulator model – is necessary. That is, the act of regulating a system in the simplest optimal manner really only demands that the actual regulator *device* be a model of the system. It makes no comment whatsoever about any technical specifications we might use to describe the details of this representational relationship. This is a subtle point, but unless we recognize it we might misinterpret the theorem to be saying that even the technical specifications are required, but this is really a misinterpretation of the theorem. The only reason these technical specifications even show up in the statement of the theorem is so that we can talk about the theorem. Remember, the mapping we called $k$ that makes the mapping $h : S \to R$ a model of the regulator/system ensemble *is entirely up in our own heads*. It really has nothing to do with the theorem, unless we are trying to understand it. As far as these good-regulators are concerned, whatever desire we may have to talk about them or to prove theorems about them is irrelevant to the actual task of regulation. As soon as we stop all the discussion, the need for the technical specifications appears to vanish. On the other hand, technical specifications notwithstanding, the good-regulator device itself absolutely *must be* a model of the system it regulates. That is what C&A's theorem establishes.

As these observations relate to our dedicated father, the Good-Regulator Theorem only proves that the father must act *as if* his behavior were being guided by the assembly instructions. It does not prove that the actual assembly instructions themselves are necessary to the successful assembly of the table. Of course, having observed all of this, it does appear to be a matter of *empirical fact* that such technical specifications are also necessary, or at least extremely useful. A disciplined glance throughout the world of human behavior reveals that most of what makes modern civilization "modern" would be practically impossible without these sorts of technical specifications. Ping pong table assembly instructions, architectural blueprints, computer software specifications, library catalogs, cooking recipes, road maps, grocery lists, income tax returns, contracts, etc. – these are all examples of control-models or technical specifications that we humans use to help ourselves become good-regulator models of the systems we need to regulate and it is hard to imagine how we could ever hope to manage all of that complexity without these kinds of artifacts.

But it's important to acknowledge here that the Good-Regulator Theorem does not prove that *any* of these sorts of artifacts are necessary to accomplish the regulation that we wish to accomplish. The only thing it proves is that to the extent that we wish to manage all of that complexity in the simplest, optimal manner possible, then we have to become (or find or build) good-regulator models of those systems. If it should turn out that we do happen to need these artifacts in order to become (find or build) these good-regulator models, well, then that's what we had better do, but this is really an empirical question. We really can't use C&A's theorem to argue that we must have, say, assembly instructions, blueprints or grocery lists. The best we could say is that the theorem proves that we must act *as-if* we had these artifacts. This in itself is a fairly strong statement, but to the extent that we can act this way without them then clearly we don't actually need them.

The upshot here is that when Conant and Ashby write that "the would-be model-maker now has a rigorous theorem to justify his work" it is important to recognize that it isn't simply the Good-Regulator theorem that justifies the model-maker's work. The fact is that most real-world model makers are primarily concerned with the development of technical specifications that are then used to build these good-regulator models, and as we have seen, the theorem does not prove that these technical specifications are really necessary. What does justify the model-maker's work is actually an inductive argument that begins with the Good-Regulator Theorem and then adds the empirical fact that without some sort of technical specification to guide the construction of the good-regulator model, the whole process is either doomed to fail completely or at least be very awkward and expensive.

This is, I believe, what Conant and Ashby mean when they write "whether it was done explicitly, or simply developed as the regulator was improved." Somehow or another the good-regulator must become a model of the system it regulates. Although it is a legitimate possibility that this could just happen as the result of some massive binary coincidence, and although it might also occur as the result of a protracted blind and incremental struggle towards successful regulation, as it concerns the types of

real-world applications to which Conant and Ashby were referring, the whole point of model-making is to bring about this regulatory representational relationship between the system and its regulator *explicitly.* This means that *first* the technical specifications are developed, and *then* the good-regulator is actually built. Of course, such an ordering does not mean that the whole process can't also be iterative and require many micro-cycles of this plan-first-then-build approach, but it does imply that the planning (modeling) part is a crucial part of the process.

The last observation I would like to make has to do with the fact that the Good-Regulator Theorem is really just a statement about a very special type of regulator: one that finds the absolute minimum output entropy that is possible to attain under the circumstances and which attains this minimum entropy in the simplest possible manner. This is actually a very strict standard and when we consider that many real world situations could involve both systems and regulators rather large behavioral repertoires, it becomes apparent that the search space we must look through in order to find these highly idealized good-regulators is astronomically huge. This suggests that it could be very unlikely that the types of regulators we encounter on a daily basis come anywhere close to these idealized entropy-minimizing good-regulator models. To the extent that this is true, we might wonder about the applicability of C&A's theorem to all of these possibly sub-optimal regulators. Do they also have to be models of the systems they regulate?

As it turns out, the answer to this question is "yes, they do". The reason they do is that C&A's theorem is actually a special case of a much more general theorem which can be paraphrased as "Even decent regulators must be models of the systems they regulate." And although the proof of this assertion is actually much, much simpler to follow than the proof we have studied in this essay, it also falls outside the scope of this essay and so I will now refer you to this essay's sequel in which this proof is given along with a number of other results[8].

[8]See the essay "Every Good Key Must Be A Model Of The Lock It Opens" available on-line on the "Educational Materials" page at www.goodregulatorproject.org.