A Triviality Worry for the Internal Model Principle

Imran Thobani

Abstract

The Good Regulator Theorem and the Internal Model Principle are sometimes cited as mathematical proofs that an agent needs an internal model of the world in order to have an optimal policy. However, these principles rely on a definition of "internal model" that is far too permissive, applying even to cases of systems that do not use an internal model. As a result, these principles do not provide evidence (let alone a proof) that internal models are necessary. The paper also diagnoses what is missing in the GRT and IMP definitions of internal model, which is that models need to make predictions that *represent* variables in the target system (and these representations need to be usable by an agent so as to guide behavior).

Keywords: internal model; representations; neural networks ; optimal policy; triviality worry; neuroscience; artificial intelligence

1 Introduction

The idea that an agent needs an internal model of the world in order to behave intelligently or optimally seems intuitive. Indeed, this idea has a long history (Craik, 1952) and seems to motivate many research programs across the cognitive sciences, such as internal models in neuroscience and model-based AI. In fact, some researchers have gone so far as to claim that we can prove the necessity of internal models under very general assumptions about the agent and task. The first such "proof" was called the Good Regulator Theorem (Conant & Ashby, 1970), and later, more detailed elaborations of this theorem came out in the field of control theory under the name of the "Internal Model Principle" (e.g. W. Wonham (2018)). In turn, various researchers in cognitive science have cited both principles as supporting the importance of internal models for optimal policies – support that ostensibly carries the force of a mathematical theorem, rather than mere intuitions.

Given that the Good Regulator Theorem (GRT) and Internal Model Principle (IMP) are sometimes cited as support for the necessity of internal models for optimal policies or optimal control, it is worth examining them to see whether they do support such a claim. I will argue that reliance on the GRT and IMP turns out to be a mistake, because both principles are vulnerable to a triviality worry. By this, I mean that they employ a definition of internal model that is overly permissive, and so do not tell us anything about internal models properly construed (i.e. models worth caring about). I will then discuss what is missing from the overly permissive definition of model, namely, that the model should make predictions that *represent* predicted variables in the target system, in a way that helps explain an agent's behavioral success on some task.

2 Background on the Good Regulator Theorem and the Internal Model Principle

Both the GRT and IMP purport to show that under certain assumptions, a good regulator must contain an internal model of its environment (the exosystem). Both of these principles have come up in recent discussions in cognitive science, often being cited as "proof" or at least strong evidence that a good controller of a system must contain an internal model of that system. Here are a few recent examples from the literature.

(Piantadosi, 2021, p. 8) cites Conant and Ashby's Good Regulator Theorem as showing that in order for a system to be a good regulator (controller) of another system, the former must contain a model of the latter. The notion of model is cashed out as an "isomorphism," meaning that the dynamics of the model in some way mirrors the dynamics of the target system. As I will argue later in this paper, I think the notion of isomorphism (which is roughly what the IMP model criteria capture), while relevant, isn't on its own strong enough to do justice to what we want from models (either cognitive or scientific models).

In a perspective paper, Seligman, Railton, Baumeister, and Sripada (2013) likewise appeal to the GRT as a proof that internal models are necessary for "the brain to be a good regulator of interactions with the environment" (p. 124). It is important to note that what the authors have in mind by "model" is some internal representation that could support prospection (forming expectations about future events and simulating possible futures). As will become clear, this is a stronger notion of model than the one that appears in either the GRT or IMP.[1]

Reliance on the GRT or IMP is not limited to psychology and neuroscience, but also comes up, unsurprisingly, in robotics, where one would expect connections to be made to work in control theory such as the Internal Model

[1]Some other examples of works that cite the GRT or IMP (or in many cases, both) from neuroscience or psychology are Barrett (2017), Cheung (2020), Graziano, Guterstam, Bio, and Wilterson (2020), Moray (1999), and Seth (2015).

Principle. For example, a paper providing advice to incoming robotics graduate students cites both the GRT and IMP as showing that learning an optimal policy entails learning a model (Atkeson, 2020, p. 8).[2]

There has been some limited discussion of the GRT (e.g. Baez (2016)), but no detailed treatments. The IMP is presented as a more detailed and rigorous version of the GRT, but there has been no discussion of the IMP (apart from citing it). Therefore, I discuss both the GRT and IMP in this paper. In what follows, I briefly introduce these principles by stating the assumptions of the theorems, and most importantly, the criteria they use for an "internal model." This sets the stage for my main argument, which is that the criteria employed by GRT and IMP for an "internal model" are vulnerable to a triviality worry.

3 Triviality Arguments

Before discussing the GRT and IMP in detail, it may be helpful to discuss briefly what a triviality argument is. A triviality argument attacks a proposed definition of a concept (like computation, internal model, or representation). It purports to show that under the proposed definition, the concept becomes overly permissive. That means the definition applies not only to all the cases that we think the concept should intuitively apply to, but also many cases that it should not apply to.

In general, triviality worries leave us with the following options: reject the proposed definition (and come up with a better one in the future), reject the concept altogether (as it no longer does the work it was supposed to do), or bite the bullet (realize that the concept is far more permissive than intuition or established wisdom would lead us to believe, and accept whatever theoretical implications follow).

The problem with both the GRT and IMP is that they employ a definition of "internal model" that is vulnerable to this kind of worry. To argue for this claim, I will describe a system that does not count as an internal model yet does satisfy the model criteria that the IMP and GRT rely on, thus showing these criteria to be overly permissive. Also, because (as I will argue later) we have reason to think that there is a stronger definition of internal model that does not trivialize, there is good reason to reject the definition of internal model that is employed by the GRT/IMP in favor of the stronger notion of internal model.

4 The Good Regulator Theorem and the Internal Model Principle

The goal of this section is to briefly introduce the GRT (Conant & Ashby, 1970) and the Internal Model Principle (W. Wonham, 2018). While there are other problems with the proof of the GRT, the emphasis will be on the criteria for internal models that these works rely on, which I argue in the next section

[2] Another example from robotics is Roberts, Koditschek, and Miracchi (2020).

are overly permissive. This generates a triviality worry for the IMP model criteria (as well as the GRT model criteria, which are weaker than the IMP model criteria). In addition, I will discuss the assumptions that the IMP makes to clarify its scope (the kinds of systems it applies to).

To start with, here is what the GRT says. Suppose R is a variable representing the state of a physical system, a regulator, which is controlling another system whose state is represented by variable S. Z represents the value of some outcome that is to be regulated, and we imagine that Z is determined by S and R jointly, as in the following diagram copied from (Conant & Ashby, 1970, p. 90).

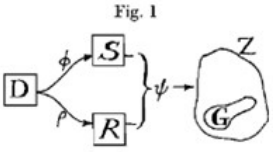

Fig. 1 The set up for the Good Regulator Theorem.

Good Regulator Theorem (GRT). *If system R is a good regulator of system S (with respect to outcome Z) and subject to this condition, the entropy $H(R \mid S)$ is minimized, then R is a deterministic function of S.*

The authors consider various possible definitions of what it means to be a "good regulator." The most intuitive one is that there is a subset G of possible outcomes that count as "good" (see Fig. 1) and that a good regulator is one that ensures that $Z \in G$. However, the definition of "good regulation" that is actually used in the GRT proof is different: the authors say that R is a "good regulator" if it minimizes the entropy $H(Z)$. It's not immediately obvious what the entropy-minimization condition has to do with the subset-G condition, but one possible interpretation is as follows. In many practical situations, the choice of G will be such that keeping Z in a very narrowly defined subset G will require keeping Z as tightly controlled as possible, or minimizing its entropy. So, while the definitions are not obviously equivalent, for some practical choices of G, keeping Z in G will require minimizing $H(Z)$.[3]

The word "model" doesn't occur in the above statement of the GRT, but the idea seems to be that R is a "model" of S if R is a deterministic function of S. This is a weak notion of model, because the mapping from S to R could be any mapping and could lose a great deal of information about S if the mapping is non-injective (imagine that every value of S gets mapped to the same value of R).

[3]Of course, minimizing $H(Z)$ could also be achieved by forcing Z to be some particular *bad* target value. As a reviewer has helpfully pointed out, it is possible that Conant and Ashby believed that confining Z to a narrow regime is the hard part, and where that regime should lie is more a matter of choice.

To illustrate this point, we can imagine a case of "regulation" involving putting a golf ball into the hole. Z can potentially take on two values, "in the hole" or "not in the hole," S is the location of the hole, and R is the angle of the putter, which can be between 60 and 120 degrees (so the putter always hits forward, not backward or to the side). But as it turns out, the ground is always sloped so that the ball will end up in the hole regardless of the angle of the putter. In that case, the "simplest regulator" (the one minimizing the entropy of R) will be fixed to some arbitrary value of R, say, 63 degrees. It doesn't matter what angle R is, since the ball always ends up in the hole. This is a good regulator (because any regulator in this setup is a good regulator), and R is a deterministic function of the location of the hole (because R is fixed to be a constant value). However, there's no interesting sense in which R models the location of the hole. In fact, because R is set to be a *constant* value, R doesn't seem to be a model of anything (neither the location of the hole, nor any other states such as whether or not the ball is in the hole).[4]

In the 1970s, more rigorous versions of the Good Regulator Theorem in control theory were developed that went under the umbrella term "Internal Model Principle" (which is more like a family of related principles). I will focus on one recent version of the Internal Model Principle by W. Wonham (2018). The reason for picking this version is that it is the most general one – unlike the versions of IMP that applied to linear systems only, this version applies to any system that satisfies the assumptions laid out below.

The setup (a common setup in control theory) is that we have a physical system consisting of three parts - an exosystem, a controller, and a plant. We can denote the entire system state by S, and denote the states of the individual components (exosystem, controller, and plant) with the symbols E, C, and P, respectively. The exosystem can be thought of as a source of disturbances to the plant, which the controller is supposed to respond to optimally (for example, by canceling out those disturbances so the plant remains in some desirable target state).

In general, the exosystem state evolves over time according to a well-defined transition function which maps the exosystem state at one timepoint to the exosystem state at the next timepoint. Moreover, the exosystem state at a given time may depend on some parameter μ (W. Wonham, 2018). Changing

[4]There are further issues with the GRT proof besides its notion of a "model." In brief, here is how the GRT proof works. In order to minimize $H(Z)$, the regulator R must behave so that Z is fully determined by S (in other words, once you fix S to a specific value s, there's no remaining uncertainty about Z). Now suppose that R is a good regulator. For a specific value $S = s$, R could potentially take on multiple different values r with positive conditional probability, but it doesn't matter which of these values R has, because they all result in the same value of Z. So for each s, you could just force R to take on one of those possible values, i.e. you could design R so that it's a deterministic function of S and it would still be a good regulator. Conant and Ashby thus conclude that the simplest possible regulator R will be a deterministic function of S ("simplest" in the information theoretic sense of minimizing $H(R \mid S)$).
Note that this proof only shows that, if there exists a good regulator of a system, then there is *some* good regulator R that is a deterministic function of S. It does not prove that every good regulator must be a deterministic function of S. Furthermore, it's not clear that a good regulator R that minimizes $H(R \mid S)$ really is the simplest one from a design standpoint. After all, it could take more work to design a good regulator whose states are a non-noisy function of S, compared to one that has some noise.

μ does not change the transition function over exosystem states, but it does change the initial state of the exosystem (at $t = 0$) and therefore also the subsequent states. In general, we can denote the state of the exosystem at a given time as $E_\mu(t)$, reflecting the dependence of the exosystem state on both t and μ.

Here are the four assumptions of the IMP:

1. Internal stability: The states of the controller and plant are fully determined by (that is, are a function of), the state of the exosystem. In other words, you could predict the states of the controller and the plant based on the state of the exosystem, assuming you knew the relevant function from exosystem state to controller/plant states.[5]
2. Exosystem detectability: The exosystem is observable by the controller as long as regulation is perfect. This means that whenever regulation is perfect, it is in principle possible to determine the initial state of the exosystem based on the full sequence of controller states over time (W.M. Wonham, 1976, p. 737).
3. Perfect regulation: The system (at least asymptotically) remains in a target subset K of good states.
4. Error feedback: The controller is autonomous whenever the system is in a good state. Controller autonomy means that the next controller state is a function of the current controller state. Presumably, this condition is referred to as "error feedback" because the controller state is only affected by the exosystem state through an error feedback signal. That is, whenever the controller is *not* in a good state, it gets an error signal (caused by the signal coming from the exosystem), to correct its behavior and return the system to a good state. As long as the system state is good, there is no error, and thus no correction needed, so the controller continues on a well-defined predictable trajectory.

Before continuing, it is worth commenting on some of the motivations for these assumptions, particularly exosystem detectability and error feedback, as these conditions play an important role in the IMP proof. Exosystem detectability is important so that the controller has some way of tracking the state of the exosystem over time. This is important because the exosystem may have different initial states, as mentioned above, depending on its parameter μ. If the trajectory of controller states cannot predict the initial state of the exosystem, then it is hard to see how the controller states could be used to identify subsequent states of the exosystem either (i.e. how one could find a mapping from controller states to exosystem states).

Controller autonomy (which comes up in the error feedback condition) is important because the controller is supposed to instantiate a genuine *model* of the exosystem, which is not dependent on getting information at each moment

[5]It is worth mentioning that the internal stability assumption is closely related to the GRT model criterion, since part of what it says is that the controller state is a function of the exosystem state.

from the outside world to decide what to do. Thus, any information the controller has about the exosystem must be in the controller itself. Of course, the controller may have received this information in the past, before it became autonomous, but once it is autonomous, it is not gaining any new information from the outside world.

Exosystem detectability and error feedback may seem to be in tension with each other. On the one hand, exosystem detectability suggests that the controller states are causally dependent on the exosystem (at least to the extent that they are sensitive to its initial state). On the other hand, error feedback suggests that the controller is causally insulated from the exosystem once regulation is perfect. However, these two conditions are perfectly compatible. At early time steps, non-zero error signals can provide the controller with enough information to establish exosystem detectability. Eventually, though, the error signals converge to 0, and the error feedback condition says that under *perfect* regulation, the controller must be autonomous.

Given the above assumptions, the Internal Model Principle is stated as follows: For any system S satisfying assumptions 1-4, the controller system C instantiates an "internal model" in the sense that the following two conditions hold (i.e. these are the criteria the IMP uses for an "internal model"):

IMP Criteria for Internal Model

1. Controller autonomy: There is a unique transition function from $X_C \to X_C$ (the controller states) that describes the dynamics of C (note that this is essentially the error feedback condition).
2. Injective mapping: The controller dynamics "mirror" the dynamics of the exosystem in the sense that there is an injective function f such that $C = f(E)$. Injectivity ensures that the dynamics of the exosystem are reflected in the dynamics of the controller (so the controller can't just be in a constant state, for example, if the exosystem is non-constant). While the internal stability assumption already guaranteed that there is a function from exosystem states to controller states, the IMP model criteria further require that that function be injective.

Controller autonomy, the first condition of the IMP model criteria, follows trivially from error feedback and perfect regulation. So what the IMP is actually *proving* is the second condition, that the mapping from exosystem states to controller states (the existence of which was already assumed by the internal stability condition) must furthermore be *injective*.

It is worth commenting on why one might have expected the IMP model criteria to provide a good definition of an internal model (even though I will eventually argue that they do not). The IMP model has an "autonomous" dynamics that can be described with a transition function over states. The dynamics of the controller's transition function mirror the dynamics of the exosystem, which suggests that the controller has in some sense replicated the exosystem dynamics. This seems to count as a kind of structural resemblance, which is often considered to be an important feature of models.

The IMP notion of model is somewhat stronger than that in the Good Regulator Theorem. First, note that the GRT model criterion just says that the regulator state is a function of the exosystem state, and this criterion is entailed by the IMP criteria (which say that there is a function from the exosystem state to the state of the controller). However, the IMP criteria go beyond the GRT criterion in a couple ways. First, the function from the exosystem state to the controller state is injective, whereas in the GRT the function need not be. Also, we know that the controller dynamics is "autonomous," as described above, and this condition does not come up at all in the GRT.

Because the IMP model criteria are stronger than the GRT criteria, we can focus on the IMP criteria from now on. If a triviality worry shows the IMP criteria to be overly permissive, then any logically weaker criteria like the GRT criteria must also be overly permissive.

5 The Triviality Worry: A Counterexample to the IMP Model Criteria

To show that the IMP model criteria are vulnerable to a triviality worry, it would suffice to give a clear counterexample. The counterexample would have to be a case of a system that we know does not instantiate an internal model, yet does satisfy the IMP model criteria.

The counterexample is as follows. Imagine a suspension-equipped bicycle riding over bumpy terrain. In this example, I take the exosystem to be the terrain under the front bike wheel at any given moment, the plant to be the bike frame, and the controller to be a suspension spring.

Let us start by defining the exosystem and plant states. I define the exosystem state E as being an ordered pair $(H, \Delta H)$[6], where H is the terrain height under the front wheel, and ΔH is the *change* in terrain height from that time to the next. When the terrain is level at that point in time, $\Delta H = 0$. The "plant" state is the change in displacement of the bike frame, ΔD, which will be 0 if and only if the terrain is level ($\Delta H = 0$). I will stipulate that regulation is considered "perfect" (i.e. the system is in a good state) whenever the bike is level ($\Delta D = 0$, which exactly corresponds to whenever the terrain is level, or $\Delta H = 0$).

Now I will define the controller states in such a way that the IMP model criteria are satisfied. I'll start with the second criterion and come back to the first one after that. Define the "controller" state C as an ordered pair $(L, \Delta L)$, the length of the spring and how much the length changes at that time point. The spring length will be an injective function of the amount of upward force on the spring, which will be an injective function of the height of the terrain at that point. This means that the controller state is going to be an injective function of the exosystem state. Thus, the second IMP criterion is satisfied.

[6] The reason for including ΔH as part of the exosystem state is that it will allow me to define an "error signal" in terms of changes in force on the spring that drives the controller state. Regulation is perfect when this error signal is 0 (which occurs when the terrain is level). This allows me to later argue in this section that the IMP's error feedback assumption is satisfied in this situation.

In order to secure the first IMP criterion, there must be a well-defined transition function over controller states. One way to ensure this is to stipulate a transition function for exosystem states (terrain heights and changes in height) over time. Here is one such transition function α_E for the exosystem. First, to get the next height based on the previous terrain height, we just use the definition of ΔH, which we included in the exosystem state:

$$H(t+1) = H(t) + \Delta H(t)$$

Second, we stipulate a transition rule for $\Delta H(t+1)$ as:

$$\Delta H(t+1) = \Delta H(t)/2 \text{ if } \mid \Delta H(t) \mid \geq 0.1$$

$$\Delta H(t+1) = 0 \text{ otherwise}$$

Eventually, $\Delta H(t) = 0$ (since the above rule just makes ΔH smaller and smaller until it is below some threshold, e.g. 0.1, and then becomes 0 thereafter). That means that the system will eventually be in a "good" state, and perfect regulation will hold. Now that we defined some transition function for the exosystem, and we know that the controller state is an injective function of the exosystem state, we guarantee that there is a transition function for the controller state. This secures the first criterion of the IMP model criteria. Now we have satisfied both IMP criteria.

It's also worth noting that in addition to the two IMP model criteria (controller autonomy and injective function from exosystem state), the four IMP assumptions also hold of the bicycle-terrain system. Internal stability is clearly secured because the controller and plant states are a function of the exoystem state. Exosystem detectability holds because you can infer the initial height of the terrain (and its initial change in height) based on the full sequence of controller states. We already saw that perfect regulation holds asymptotically. Error feedback holds because we can define the "error signal" that externally drives the controller to be the change in force on the controller at a given time, which will be non-zero if and only if the terrain is not level at that time. So when regulation is perfect, the error signal to the controller will be 0, which means the length of the spring isn't changing because the terrain height isn't changing.

Despite satisfying the IMP model criteria, the bike suspension isn't a model of the terrain height in an intuitive sense. There are at least two reasons for this.

The first is that the controller (the bike suspension system) doesn't seem to *use* the "model" of the terrain ("model" in the IMP sense, i.e. the controller's dynamics mirroring the exosystem dynamics via an injective mapping). It's not that the controller *uses* the injective mapping to predict the terrain heights or figure out what state it should be in. Rather, the mirroring of controller dynamics and exosystem dynamics is a byproduct of the fact that the terrain height directly affects the spring length. If we assume that internal models

are supposed to *represent* the dynamics of the target system, and that representations should actually be used by an agent, then the suspension system doesn't seem to have a real internal model, because it's not using any kind of representation of the terrain.

Perhaps a *person* could use the spring length as a representation of the terrain height. Assuming we *know* the injective function from spring length at time t to terrain height at time t, and we know the state of the controller at a given time t, we could then infer the terrain height at that same time by using the injective function from terrain height to spring length. Then it is plausible that *we* could use the spring length to make predictions about the terrain height. Would the spring length then count as a genuine model of the terrain height?

Even then, the controller would not be a model. The reason is that you cannot use the suspension to *simulate* how the terrain height beneath the bike wheels will change over time as the bike moves over the terrain. There is a sense in which the suspension spring has information about the terrain height, since if you know the injective mapping from the spring length to the terrain height, you can predict what the current height must be based on the current spring length. However, you cannot run the controller offline: it only has information about the environment as long as it is being causally impacted by the terrain at each time-step. Insofar as the controller needs to be causally affected by the current terrain height in order to contain any useful information about it, it's hard to see how the controller is playing the role of a *model*, which should stand in for the terrain in the sense of allowing us to make predictions about the terrain even when the model *isn't* getting information at each moment from the terrain. The problem here is that all the controller's information about the environment is coming, moment by moment, from the environment (via its causal effect on the controller), rather than being a robust feature of the controller itself. The fact that the suspension spring doesn't instantiate a genuine model yet does satisfy the IMP criteria demonstrates that the IMP criteria are too weak, which is the central point of the paper.

At this point, I have argued that the bicycle example can be described in a way that meets the technical conditions of the IMP, even though it is not an example of a genuine model. However, perhaps one could object that the IMP could be amended to rule out this sort of example. Specifically, why not strengthen the error feedback assumption and the controller autonomy condition to require that the model states are not causally affected by the exosystem states during perfect regulation (that is, if you took the controller offline, it would still mirror the exosystem states via the injective mapping)? Once the IMP conditions are appropriately strengthened in this way, they would no longer apply to the bicycle example.

There are two responses I have to this objection. First, the GRT and IMP are generally invoked to motivate the need for internal models *without* specific attention being paid to whether or not the controller states are causally decoupled from the exosystem in the way just described. For example, in a

perspective piece, the authors write that "for the brain to be a good regulator of interactions with the environment, both physical and social, it must build and use a model of that environment" (Seligman et al., 2013, 94). So, if we did strengthen the IMP conditions to require causal insulation of the controller from the exosystem, those who want to rely on the IMP would then need to assess whether these strengthened conditions hold for the cases in which they think a given controller *must* instantiate an internal model.

A more serious issue is that, even if we require the controller to be causally decoupled from the exosystem during perfect regulation, the IMP is *still* vulnerable to a triviality worry. The next section explains why.

6 A More Serious Triviality Worry for the IMP

In this section, I will argue that, even if we add a requirement to the IMP that (under perfect regulation) the exosystem states must not causally affect the controller states, the IMP model criteria will still apply to systems that do not instantiate genuine models. Specifically, I will show that the IMP model criteria apply to a class of systems that satisfy two conditions. Because these conditions are fairly minimal, this class contains many counter-examples to the Internal Model Principle. I will give one such counter-example at the end of this section.

Here are the two conditions.

1. The system must contain a *clock* as defined by Chalmers (1996). A clock is a component of the system that reliably transits through a sequence of states $i(t)$, and crucially is in a different state at every point in time.
2. The system must have an "input component" that is sensitive to the *initial* state of the exosystem $E(0)$ but not sensitive to subsequent exosystem states. Formally, the state j of this component is an injective function of $E(0)$, i.e. $j = \phi(E_\mu(0))$ where ϕ is an injective function from the possible exosystem states to controller states. As a result, j must be constant over time (but it does depend on the exosystem parameter μ which determines the exosystem's initial state).

   This second condition is included to secure exosystem detectability, at least in the case where the initial state of the exosystem is not fixed and could be one of many possible states (determined by the parameter μ).

If a system satisfies the above two conditions, then we can describe its state at time t as an ordered pair $(i(t), j)$, ignoring any other details about the system. For convenience in relating this discussion to the IMP setup, I will refer to a system that satisfies the above conditions as the "controller," although whether the so-called "controller" is actually doing anything interesting plays no role in my argument that it will satisfy the IMP model criteria.

Note that the state $(i(t), j)$ is not causally affected by the exosystem state after $t = 0$. The clock $i(t)$ follows its own reliable sequence. And j is set by

the initial exosystem state, but it isn't sensitive to the subsequent exosystem states (for example, if j is a mark that is inscribed on a rock at $t = 0$).

The crucial move in my argument will be to *coarse-grain* the controller states, i.e. to form equivalence classes of the ordered pairs, such that the exosystem states can be mapped injectively to the coarse-grained controller states. This move is very similar to moves made in the literature on triviality worries for computational implementation, especially Chalmers (1996) (see Sprevak (2018) for a helpful overview).

Let $E_\mu(t)$ denote the state of the exosystem at t, given parameter μ. The corresponding controller state is then: $C_\mu(t) = (i(t), j)$, where $j = \phi(E_\mu(0))$, with ϕ injective. The clock component $i(t)$ does not depend on μ because the clock *reliably* transitions through its sequence of states regardless of how the exosystem is initialized. Intuitively, one can think of the clock $i(t)$ as "encoding" (loosely speaking) the time, t, while j "encodes" the initial state of the exosystem. This means that, if one knows the state $C_\mu(t)$, one can in principle determine the initial state of the exosystem $E_\mu(0)$ as well as the time t at which $C_\mu(t)$ occurred. One can then in principle determine $E_\mu(t)$ (the exosystem state that occurred at the same time as $C_\mu(t)$) by repeatedly applying the exosystem transition function to the initial state $E_\mu(0)$, t times.

Given these facts, we can define the following equivalence relation between controller states: $C_\mu(t) \sim C_{\mu'}(t')$ if and only if $E_\mu(t) = E_{\mu'}(t')$. Informally, this means two controller states are considered equivalent if and only if the corresponding exosystem states are identical. For a given controller state $C_\mu(t)$, let $C^*_\mu(t)$ denote the induced equivalence class containing $C_\mu(t)$: $C^*_\mu(t) = \{C_{\mu'}(t') \mid E_\mu(t) = E_{\mu'}(t')\}$. Each equivalence class is a coarse-grained controller state.

To show that the IMP criteria are satisfied, we now need only show that there is an injective function from the exosystem state $E_\mu(t)$ to the coarse-grained controller state $C^*_\mu(t)$. Let $f(E_\mu(t)) = \{C_{\mu'}(t') \mid E_\mu(t) = E_{\mu'}(t')\} = C^*_\mu(t)$. That is, we pair each exosystem state with the set of all controller states that ever co-occur with that exosystem state. Note that f is injective because, if $C^*_{\mu'}(t') = C^*_\mu(t)$, then by definition of the equivalence classes, $E_\mu(t) = E_{\mu'}(t')$.

Now we have shown that there is an injective function from the exosystem states to the coarse-grained controller states. Controller autonomy is also established because the exosystem was assumed to have a well-defined transition function, and with the controller states being an injective function of exosystem states, the controller also now has a well-defined transition function. So we have established that the IMP criteria hold of any controller satisfying conditions 1 and 2, provided that the exosystem in question has a well-defined transition function.

Despite satisfying the IMP criteria, satisfying conditions 1 and 2 above should not (in general) mean that the controller instantiates a *model* of the exosystem in any intuitive sense.[7] To illustrate this point, let us imagine a

[7] Of course, it may be the case that some systems *do* satisfy the conditions discussed here and *also* instantiate genuine internal models. The point is just that conditions 1 and 2 are not sufficient for a system to instantiate a genuine internal model.

rock sitting on the beach, which will be our "controller." The exosystem is the waves hitting the beach, which hit the beach at regular times $t = 0, 1, 2$, and so on. The state of the exosystem is the height of the wave hitting the beach. To ensure a well-defined transition function exists for the exosystem state, I will suppose that the wave height is cyclical, and indeed follows a very simple alternating pattern between two states: HIGH and LOW.

The state of the rock, following the discussion in this section, is an ordered pair of states $(i(t), j)$. Let j represent the location of the rock. To satisfy condition 1, I stipulate that the rock contains a clock, perhaps because of reliable patterns of radiation emission that it undergoes (Chalmers, 1996). To satisfy condition 2, suppose that at time $t = 0$, a wave hits the rock on the beach. If the wave is HIGH, the rock gets flung high up the beach, above the waves, where it remains from $t = 0$ onward. If the wave is LOW, the rock gets pulled down into the ocean and settles on the seabed, where it remains from $t = 0$ onward. So j will be in one of two states - BEACH or SEABED - depending on the initial height of the waves hitting the beach and will be constant over time starting at $t = 0$.

Because the rock satisfies these two conditions, it is possible to coarse-grain the states of the rock in such a way that there exists an injective mapping from the cyclical states of the waves to the coarse-grained states of the rock. That is, the rock satisfies the IMP criteria for an internal model. However, intuitively, the rock does not instantiate a genuine model of the cyclical behavior of the waves.

One reason the rock does not instantiate a genuine internal model is that the rock does not *use* the "model" to do anything. Here, the "mirroring" of the rock's dynamics and wave dynamics is just a consequence of the fact that the rock contains a clock, the fact that the rock's location from $t = 0$ onward is an injective function of the initial height of the waves, and the fact that the states of the rock (construed as ordered pairs of clock states and rock locations) can be suitably coarse-grained such that we can find a mapping between the wave heights and the coarse-grained states. If we assume that internal models are supposed to *represent* the dynamics of the target system, and that representations should actually be used by an agent, then the rock doesn't seem to have a real internal model, because it's not using any kind of representation of the waves.

As in the previous section, one could ask whether a *person* could use the rock's coarse-grained states as a representation of the wave height at any given time. Assuming we know the injective mapping from wave heights to coarse-grained rock states, and we know the coarse-grained state of the rock at a given time t, we could then infer the wave height at that same time (using the fact that the mapping is injective). Then it is plausible that *we* could use the rock to make predictions about the wave height. Would the rock then instantiate a genuine model of the terrain height?

Using the rock as a model in this way would require some independent way of identifying the coarse-grained states. As I have defined them, the coarse-grained states of the rock are defined as equivalence classes of controller states that correspond to the same wave height. If this is the *only* way to pick out these states, then we would have to refer to the exosystem state in order to know which of the coarse-grained states the rock is in at any given time, and then we wouldn't be using the rock to make *predictions* about the exosystem states. To use the rock as a model of the wave heights, we would need some independent way of identifying the coarse-grained rock states (e.g. on the basis of measurable properties of the rock) that did not require already knowing the height of the waves.

The most obvious way to do this would be to just use a list of the states that figure into each equivalence class, and then check (either manually, or using a computer or other man-made device) whether the rock's non-coarse-grained state $(i(t), j)$ is in a given list or not. But the clock has infinitely many states, so the list of states for each equivalence class would be infinitely long. It will not be physically possible to decode the coarse-grained state of the rock in this way, because you would need to store an infinitely long list of states. Therefore, in order to decode the coarse-grained state of the rock at any given time, there would have to be some *genuine similarity* amongst the states in each equivalence class (besides corresponding to the same wave height, and besides being a member of the same disjunctive set), by which we could identify the coarse-grained rock states. Since there is no reason to suppose that this will *generally* be the case, the rock remains as an intuitive counterexample to the model criteria used by the Internal Model Principle.

7 What's Missing from the IMP Model Criteria?

The preceding discussion naturally raises a further important question: how might the IMP criteria be strengthened into a more substantive notion of model? Although the main point of the paper (the triviality worry for the IMP) has already been made, in this section, I will end by offering an initial suggestion for what is missing from the IMP model criteria (while leaving a deeper investigation of these issues to future work). I suggest that in order for a system to instantiate a genuine model, variables in the model should be *used* by an agent (e.g. a controller or a person) to make predictions about variables in the target system.

By requiring an injective mapping between exosystem variables to controller variables, the IMP model criteria capture the idea that controller variables should be able to *predict* variables in the exosystem under some mapping. However, the problem is that mappings are relatively cheap. It is possible to map variables between two systems in a way that is not relevant to explaining the behavioral success of a controller (as we see in the example of the rock in the previous section). In order for the predictions to be *usable* by an agent

or controller, the mapping from those model variables to exosystem variables must be exploitable by a controller to guide its success on some task.

This way of thinking about models implies that models involve a kind of representation. This is because models make predictions that *represent* the predicted variables in the target system. That implies that whether something is or is not a model depends partly on who or what is using it (i.e. the system or agent that is using it must use variables in the model *as* predictions about something in the world).

Enactivist approaches in cognitive science will disagree with my view that models involve representation. For example, a recent paper (Ramstead, Kirchhoff, & Friston, 2020) interprets the GRT as saying that a model of the environment is instantiated in the dynamics of the controller, including its actions over time. Given the triviality worry raised in this paper, those who use the word "model" in this way should worry that their notion of model may be too weak. If representation is *not* to be a requirement for models, then enactivists will need to think of other ways to strengthen their notion of model to avoid triviality.

It is also worth noting that many theorists who believe in the necessity of internal models seem to be motivated by the intuition that internal models allow us to simulate the environment without having to perform certain actions. For example, a paper on internal models in biological control cites Craik (1952) and writes: "an internal model allows an organism to contemplate the consequences of actions from its current state without actually committing itself to those actions" (McNamee & Wolpert, 2019, 340). This motivation for internal models is hard to reconcile with the enactivist take, in which the model is *instantiated* in the controller's actions.

In saying that models are representational, I do not make the converse claim that representations in general are always models. The claim is only that one requirement for being a genuine model is being a genuine representation (and representations must be usable by an agent). This requirement doesn't seem to be captured by the IMP model criteria.

Besides being usable representations, there may also be further requirements for internal models, such as decouplability[8], i.e. being able to be run offline or in the absence of external causal influences (Grush, 1997). The bicycle example showed that there are ways to satisfy the IMP model criteria that does not result in a decouplable model. However, as I suggested, it might be possible to strengthen the IMP by requiring that, at least during perfect regulation, the exosystem state should not causally affect the controller state. Even with this strengthening of the IMP, the model would not be fully decouplable, because the error feedback setup means that if the controller ever *did* veer off-course (i.e. no longer a perfect regulator), an error feedback signal would provide external information to the controller. That means the controller is never fully causally decoupled from the exosystem. That being said, complete decouplability is a fairly strong requirement to place on internal models, and

[8] Gładziejewski (2016) calls this "detachability."

I think weaker degrees of decouplability might be enough to call something an "internal model" (Cao & Warren, 2023). The central issue is instead that the IMP notion of model can apply even to cases of systems that do not have representations.

8 Conclusion

I have argued that the Internal Model Principle and the GRT cannot be relied upon to support the necessity of internal models because they rely on an overly permissive notion of internal model. As a result, we need more theoretical and empirical work to show exactly why and when genuine internal models are useful for cognitive agents or controller systems. This suggests an exciting future research direction – namely, to develop a new Internal Model Principle that would prove that (under certain precisely specified conditions) an agent needs a genuine internal model in order to perform certain tasks.

Acknowledgements

I would like to thank Rosa Cao, Thomas Icard, Zhengyan Chang, Jared Warren, John Krakauer, David Gottlieb, Mikayla Kelley, Daniel Wolpert, John Morrison, Samuel Lippl, and Juanhe TJ Tan for helpful feedback or discussions.

References

Aronowitz, S. (2019). Memory is a modeling system. *Mind & Language*, *34*(4), 483-502.

Atkeson, C. (2020). What advice would i give a starting graduate student interested in robot learning? *Rss workshop on robotics retrospectives.*

Baez, J.C. (2016, January). *The internal model principle.* Retrieved from https://johncarlosbaez.wordpress.com/2016/01/27/the-good-regulator-theorem/

Barrett, L.F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1-23.

Cao, R., & Warren, J. (2023). Mental representation,"standing-in-for", and internal models. *Philosophical Psychology*, 1–18.

Chalmers, D.J. (1996). Does a rock implement every finite-state automaton? *Synthese*, *108*(3), 309–333.

Cheung, J.A.e.a. (2020). Independent representations of self-motion and object location in barrel cortex output. *PLoS Biology*, *18*(11).

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese*, *101*(3), 401-431.

Conant, R.C., & Ashby, W.R. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, *1*(2), 89-97.

Craik, K.J.W. (1952). The nature of explanation. , *445*. (CUP Archive)

Francis, B.A., & Wonham, W.M. (1976). The internal model principle of control theory. *Automatica*, *12*(5), 457-465.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*, 559–582.

Graziano, M.S., Guterstam, A., Bio, B.J., Wilterson, A.I. (2020). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive Neuropsychology*, *37*(3-4), 155-172.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology*, *10*(1), 5-23.

Kirsh, D., & Maglio, P. (1992, July). Some epistemic benefits of action: Tetris, a case study. *Proceedings of the fourteenth annual conference of the cognitive science society.* Vol. 29.

McNamee, D., & Wolpert, D.M. (2019). Internal models in biological control. *Annual review of control, robotics, and autonomous systems*, *2*, 339–364.

Moray, N. (1999). Mental models in theory and practice. *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application*, *223*.

Piantadosi, S.T. (2021). The computational origin of representation. *Minds and Machines*, *31*(1), 1-58.

Ramstead, M.J., Kirchhoff, M.D., Friston, K.J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive behavior*, *28*(4), 225–239.

Roberts, S.F., Koditschek, D.E., Miracchi, L.J. (2020). Examples of gibsonian affordances in legged robotics research using an empirical, generative framework. *Frontiers in neurorobotics*, *14*.

Seligman, M.E., Railton, P., Baumeister, R.F., Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on psychological science*, *8*(2), 119-141.

Seth, A.K. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. (p. 1-24). Open MIND.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Sprevak, M. (2018). Triviality arguments about computational implementation. *The routledge handbook of the computational mind* (p. 175-191). Routledge.

Sutton, R.S., & Barto, A.G. (2018). *Reinforcement learning: An introduction* (Second ed.). The MIT Press. Retrieved from http://incompleteideas.net/book/the-book-2nd.html

Wonham, W. (2018). *The internal model principle of control theory.*

Wonham, W.M. (1976). Towards an abstract internal model principle. *IEEE Transactions on Systems, Man, and Cybernetics*, *11*, 735-740.

Wonham, W.M., & Cai, K. (2019). Algebraic preliminaries. *Supervisory control of discrete-event systems. communications and control engineering.* Retrieved from https://doi.org/10.1007/978-3-319-77452-7_1