Representation and Memory

Peter Godfrey-Smith
City University of New York

A lecture for the IHPST & DEC, Paris, May 2012

1. Introduction

One hope in philosophy has been to give a general theory of signs and representation. The resulting theories have taken many forms. One example is seen in naturalistic theories of thought and language content in the 1980s. Another example is semiotics, especially in the versions of Barthes and Levi-Strauss. In contrast, a feature of some influential recent work has been a turn away from this high level of generality. This is seen, for example, in Dan Sperber and Deirdre Wilson's 1986 book on language, *Relevance*. Sperber and Wilson argued that progress in understanding language had come about by recognizing its *sui generis* character. They saw the search for excessive generality as a trap, with semiotics as a cautionary case.

Today I will talk again about the possibility of a general theory in these areas. I think that it has recently become possible to recognize a convergence between work deriving from different approaches, resulting in real progress. What is especially promising here is a combination of a particular model plus a way of understanding its application to actual cases. The topic is timely in other ways, too. As discussed in James Gleick's 2011 book *The Information*, we live in an unusual time with respect to our relationship to information. *Whatever* we think information is, we are living in a *flood* of information with respect to its quantity and availability. The printed contents of the US Library of Congress, once used by Claude Shannon as an example of a truly vast amount of information, can now be stored on a hard drive that costs about one thousand dollars. This much is clear even though the status of this whole mode of description is still very unclear. The first half of the talk will be about the model in general, and the second will apply it to a particular case.

2. Sender-receiver models

My starting point is David Lewis' model of "conventional signaling," developed in the 1960s in his dissertation and first book, *Convention*, and intended as a reply to his

advisor W.V. Quine's skepticism about meaning and analyticity. I will modify nearly all Lewis' terminology and symbolism. Assume two agents, a *sender* and a *receiver*, with the following distinction between their roles. The sender can see the state of the world but cannot act except to produce signs; the receiver can only see the signs, but can act in a way that affects them both. Each adjusts their behaviors independently through rational choice. Under some conditions, a sender and receiver can reach an equilibrium state where the sender sends distinctive signals in each state of the world, and the receiver uses these as a guide to behavior. The equilibrium is maintained by rational choice in a context of common *knowledge* and common *interest*; the sender and receiver have, and know they have, the same preferences over acts the receiver might produce in each state of the world. In effect, the sender acts as the receiver's eyes, and the receiver acts as the sender's muscles.

Lewis' main example was the case of Paul Revere in the American revolution. Revere coordinated with the Sexton of the Old North Church in Boston to send him signals indicating the behavior of the British army, using lanterns placed in a church tower – one lantern if the British were coming by land and two if by sea.

More formally, in a case like this there is a *sender's rule*, f_S, and a *receiver's rule*, f_R. The sender's rule maps states of the world to signs; the receiver's rule maps signs to acts. The output (codomain) of f_S is the input (domain) of f_R. Any sender's rule and receiver's rule when combined (composed) yield a function from states to acts, F. It is also assumed that there is one function from states to acts, F*, that is the *preferred* function. F* assumed to be one to one, and because of common interest, it is the same for both sides. Many different combinations of f_S and f_R can give rise to F*. The function F that is realized by combining the sender's and receiver's rules may or may not be one-to-one. Information can lost at either stage – by the sender refusing to distinguish some states with signs, or the receiver refusing to distinguish some signs with acts.

The point of the model is to show how sender and receiver can settle on a combination of rules that serves their common interests, and does so in a way that, according to Lewis, gives the signs *conventional meanings*. Here is a picture of a sender-receiver configuration of this kind, with the roles of these functions illustrated:

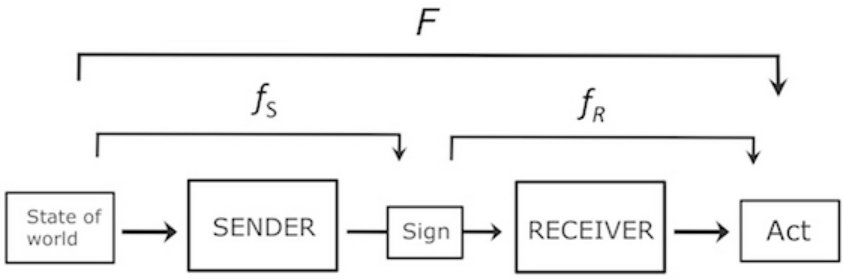

Figure 1: Sender-receiver configuration (SRC)

Lewis' discussion had little influence on naturalistic philosophy, or other philosophy aimed a "ground-up" treatment of meaning and representation, probably because he assumed rational agents with common knowledge. Then nearly 30 years later, Brian Skyrms, in a chapter of his 1996 book *Evolution of the Social Contract*, showed how to naturalize the Lewis model. Rational choice was replaced by an evolutionary selection process, and Skyrms made it clear that the agents needed to follow suitable sender's and receiver's rules can be very simple. Even bacteria sending chemical signals and evolving by differential reproduction can fit a version of the model. Since then, Skyrms has followed up his own treatment (*Signals*, 2010) and so have a number of workers including Jeff Barrett, Simon Huttegger, Rory Smead, and Kevin Zollman.[1]

The model is about signs but its focus is sender and receiver behaviors. Why does the sender keep sending? Why does the receiver pay attention to what is sent? The sender could send the same signal in every state, "pool" some states together but distinguish others, or send signs randomly. Why do one of these things rather than another? The receiver could do the same thing all the time – a cover-all behavior – could attend to some signals and ignore others, or act in a way that is a one-to-one function of what he sees. The shaping and stabilization of these rules can take place through a variety of processes, operating at different scales and degrees of cognitive sophistication. They include evolution through differential reproduction, reinforcement learning, differential imitation, and rational choice itself. In Lewis' model and Skyrms' first models, this stabilization occurred in a situation of *complete common interest*. Sender and receiver were assumed to have the same preference

[1] See for example Huttegger, Skyrms, Smead, and Zollman, "Evolutionary dynamics of Lewis signaling games..." *Synthese* 2010. See also Bill Harms' 2004 "Primitive Content...." for an early and insightful discussion of the relevance of these models to philosophy.

ordering over acts the receiver might perform in each state. That is an extreme case. Another extreme case is complete *conflict* of interest. Suppose that sender and receiver have *reversed* preference orderings over behaviors the receiver might produce in each state, and these orderings differ across states. Then if the sender sends signs that covary with the state of the world, the receiver will use them to guide actions that are as bad as possible from the sender's point of view. On the other side, if the receiver allows their acts to be sensitive to the signs, they can be exploited by the sender. So the only equilibrium can be one where the sender sends uninformative signs and the receiver is insensitive to any variation in the signs that are sent. Between the extremes of complete common interest and complete conflict of interest there are many kinds of *partial* common interest (PCI). Some of these allow the maintenance of signaling. Sender and receiver might agree on the *worst* outcome in each state, but disagree elsewhere. They might have similar orderings in some states but not others. In many cases, what results from partial common interest is partially informative signaling.[2]

The model can apply to signaling between agents (as in Lewis' Revere, and some animal alarm calls) or within a single agent. There are also cases where the model applies in a way that is clearer than the boundaries between organisms themselves, as in chemical signaling within a colonial organism or the dance of the honey-bee. Across the contexts of signaling within organisms, between organisms, and cases of uncertain boundaries, the role of common interest differs. In signaling across organisms, complete common interest will often be absent, though partial common interest of many kinds may be present. In signaling within an organism, talk of *common* interest is often inappropriate. Rather, the parts of a highly integrated system have a kind of *joint* interest that supports sender and receiver behaviors. More generally, talk of "interest" in this context is heuristic and should not be taken too seriously. In any particular context in which an SR configuration exists, there will be some mechanism of stabilization operating – evolution, learning, choice, perhaps several at once. Talk of common interest is shorthand for some set of properties that figures in a stabilization process that is relevant in that case. In a between-agent case involving evolutionary stabilization, for example, it will be an association between the reproductive outputs of sender and receiver within each

[2] See Crawford and Sobel's 1982 model of "Strategic Information Transmission" for an influential treatment of these relationships in economics, also my "Information and Influence in Sender-Receiver Models."

pairing of a receiver behavior and a state of the world. The mechanism of stabilization and the unpacking of "interest is different in other cases.

The model has another kind of generality as well. In many familiar cases, the "gap" being bridged between sender and receiver is spatial, roughly speaking – bee dances, alarm calls, Paul Revere. But sender-receiver coordination can also be used to bridge gaps in time.

In describing the sender-receiver model above I sometimes use a terminology of "information." What is the connection between the model and information theory? I will use this question to look at the historical lineage in detail.

As I see it, a range of work has been contributing pieces to the model from various different angles. Broadly speaking, the model has two origins: Lewis' model of signaling and Claude Shannon's 1948 model of information transmission. Figure 2 reproduces a famous diagram from Shannon's 1948 paper. Shannon said that *information is carried* in a set-up like this whenever the state of the signal *reduces uncertainty* about the state of the source. It is a matter of physical correlation or dependence.

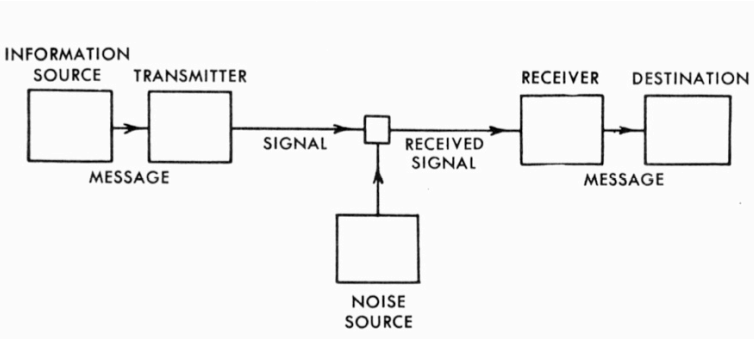

Figure 2. Shannon's diagram of a "general communication system"

Shannon drew on no philosophy, and Lewis did not seem to draw on Shannon. But the two contributions fit together in a way that can be seen in retrospect: Shannon *took for granted* the sender and receiver roles, and gave a theory of the channels that could successfully achieve coordination between them; Lewis *took for granted* the possibility of a channel, and gave a first account of how and why agents would come to play the sender and receiver roles.

Information in Shannon's sense can exist outside an SRC. Any two variables may be linked with *mutual information*, a relation whereby the value of one is

predictive of the value of the other. This relationship is seen in control systems other than SRCs, and non-evolved systems. In the sender-receiver model, the association between state and sign is a consequence of the evolution of the sender's rule, f_S, and the association between sign and act is a consequence of the evolution of the receiver's rule, f_R. It is also possible for a receiver-like agent to make use of naturally arising cues that have informational links to relevant states that are not the result of a sender.

A "second generation" is seen in philosophical work in the 1980s. Fred Dretske (1981) introduced information theory into philosophy, without considering it in the context of an SR model. Dretske's 1988 model of representation was, in effect, a model of the right hand side of an SRC, where naturally occurring informational links between an indicator and the world are responsible for the "recruitment" of that indicator as a cause of a behavior. Ruth Millikan (1984) argued that any entity that is a representation has that status as a consequence of its relations to a "producer" on one side and an "interpreter" or "consumer" on the other. She did not appeal to information, but argued that signs can "map" the world in virtue of how their producers and consumers evolved.

The goal of that 1980s work was to give an account of what semantic properties are and how they arise. As I see it, this is an outcome arising from some applications of the sender-receiver model, but it arises "along the way," as part of an account of how signing behaviors come to exist at all, and of the diverse relations that entities within control systems come to have to the world. Some applications of the model do not include a ground-up treatment of content.

Many features here hinge on the assumptions made about the capacities of sender and receiver. Starting with the simplest case, suppose that sender and receiver follow fixed and mechanical rules, and the system reaches a state where the sender's rule maps states to signs one-to-one, the receiver's rule is also one-to-one, and everyone is doing as well as they possibly could. Then it is possible to give a ground-up description of the contents of the signs; an informational analysis in the style of Dretske and a success-based analysis in the style of Millikan converge, in fact. It is not hard to come up with cases where the two modes of interpretation diverge, and in another paper I discuss the significance of this.

In a simple case with fixed sender and receiver rules, Lewis said that the signals have both indicative and imperative contents – they tell the receiver what to do as much as they tell him how the world is. Lewis called these signs "neutral."

(This is similar to Millikan's idea of a "pushmi-pullyu" representation.) Lewis also discusses a first step towards a more complex model, in a way that has bearing on the explanation of content. Sender and receiver may arrive at a situation where one side or the other acts with "discretion" – their behavior is guided not only by the state of the world (sender) or sign (receiver), but also by attention to other factors that may apply in a specific case. When there is discretion on the receiver side, but not the sender side, Lewis says the content of the sign is *indicative*, merely telling the receiver how things are. When there is discretion on the sender side, not the receiver's, the content is *imperative*, telling the receiver what to do.[3]

Why should sender and receiver benefit from opting to give one particular side discretion in this sense? Kevin Zollman argues for, and models, the idea that it will arise when one side or the other has access to additional information about the receiver's situation.[4] If so, it seems more likely that the receiver will have this extra information – as the receiver is *in situ*, where the action will take place – and hence that the receiver will have discretion and the signs will have indicative content.

A possible adaptive sequence can be seen here. Suppose a sender and receiver first reach stable signaling policies with inflexible rules on both sides. But it becomes clear that the receiver is smart enough to make use of other information, and the degree of common interest is sufficient for the sender to accept this. So a revised pattern of behavior arises, including receiver discretion. But if the receiver is smart in this way, this creates further opportunities for the sender. The sender can engage in the compression of signs, and in improvisation and innovation. The sender can make use of the receiver's ability to make *inferences*. The sender does whatever is needed to get the message across.

Sperber and Wilson, in the 1986 book I cited earlier, contrast two models of human communication. One they call the *code model:* the sender puts content into signs which is decoded by the receiver. The other is the *inferential model*: communication works by the sender providing evidence to the receiver. It can be hard to work out where the boundary is between the two models, but one contrast is clear, a contrast between use of *standardized* signs that are produced and interpreted in accordance with general rules, and *improvised* signs that make use of unique features of the immediate context and depend on novel receiver inferences. Most would agree that both phenomena are seen in human communication, and

[3] See also Sterelny's *Hostile World* on "decoupled representations."
[4] See his "Separating Directives and Assertions...." (2011, *JP*) and "Evolving Assertions and Directives" (forthcoming).

there is debate about their relative importance. A clear expression of a view emphasizing the unrepeatable and context-sensitive is seen in a recent *New York Times* article by Peter Ludlow.[5]

> [H]uman languages are one-off things that we build "on the fly" on a conversation-by-conversation basis; we can call these one-off fleeting languages *microlanguages*. Importantly, this picture rejects the idea that words are relatively stable things with fixed meanings that we come to learn. Rather, word meanings themselves are dynamic — they shift from microlanguage to microlanguage.

Highly context-sensitive sign use is outside the simplest versions of the Lewis-Skyrms model, which are concerned with the explanation of rules of sending and rules of receiving. But this is another aspect of S-R interaction, and can be linked to the simple versions of the model through the concept of "discretion" and through continuation of the same functional arguments. One difference is that when a specification of sender and receiver roles assumes that they have rich cognitive involvement with all sorts of relevant external things, there can be no ground-up naturalistic explanation of the content of signs. But as I said earlier, this is a role the model has in some cases and not others.

I think with this combination of ideas many things fall into place. Looking at the model itself, a crucial strength is that it gives a very *embedded* treatment of signs – it is all about their production and use, and the consequences of this use – but this "embedding" is handled in a minimal way, without including too much. The emphasis on use takes us away from theoretical projects that focus on alleged special properties of the signs themselves. The model steers us away from postulating any sort of "semantic glow" in signs, even if this glow is due to relations the signs have to one side or the other.

The last feature I want to emphasize in this section is a matter of how the model and the idea of "sending and receiving" are understood. As I see it, we are dealing with distinctions of degree here. There are *paradigm* sender-receiver systems and *marginal* ones, with no divider between the two. In this respect I treat semantic phenomena similarly to the way evolution by natural selection was treated in my book on Darwinism (*DPNS*, 2009). A paradigm SRC has a clear separation between the sender and receiver roles, and maintained of a pattern of sign production and use by some degree of common interest. The SRC is a natural kind, in a low-key

[5] "The Living Word," *NYT* April 23, 2012.

sense; it is something that nature builds over and over again at different scales and from different materials. But it is a kind that appears in clear versions and partial or washed-out versions.

3. Memory

In the second half of the talk I will look at a particular application of the model. The model, I said earlier, has between-agent and within-agent applications, and the bridging achieved by signs can involve either space or time. In the *psychological* case in which the bridging is across *time*, we find *memory*.

The SR model gives us a way of looking at memory that is in some ways continuous with existing work and in other ways novel. In the psychology of memory, a broadly information-processing or representational approach is standard. There are differences between a view of representation based on the SR model and a view of based on other versions of an information-processing perspective. Working within the SR model, the main idea could be expressed by saying that memory is communication; it is the sending of messages from a past self to a future self. This sounds a bit odd. To some extent that oddity is misleading, as it derives from a way of talking about the model that goes with cases with a more definite separation between S and R, distinct self-sufficient agents. Millikan's terminology "producer and consumer" might seem more natural, and I will look at another terminology later. But I do want the gap between sender and receiver to be salient in this discussion.

From viewpoint of the model, bridging gaps in time is one thing that control systems have to do. They must find ways to get present experience to bear usefully on choices that occur later. This is not easy in a biological system that is in continual flux. "Memory" is a loose term for the ways this is done. The framework I use here sees continuity between internal and external memory, between psychological memory and writing notes-to-self. I will use this analogy as a way of walking through ideas. This does not involve any claim about whether external memory is sometimes part of the user's mind, where the internal/external boundary is, as in "extended mind" debates.[6] Positions in that debate do not matter here. Biologically on-board and external memory are different resources which play complementary roles (see Clark, *Supersizing the Mind*), roles that are continually changing, perhaps especially quickly at the present time. The model can also be used to look at non-psychological forms of internal memory, in DNA and in epigenetic marks used to modify DNA.

[6] See Clark and Chalmers "The Extended Mind," 1998.

It is *possible* to see memory using the SR model. But is this is a superficial connection, or can we get real insight by taking this approach?

3.1. Common interest

Guiding questions in the model are: why does the sender send informative signs, and why does the receiver pay attention to them? Common interest, partial or complete, is important in answering these questions, though talk of "interest" is shorthand for various properties of senders and receivers that figure in a selection or stabilization process that determines sender and receiver policies.

I will treat memory here as a capacity of single organisms, not pairs or groups. In any within-organism case of signaling, we expect a lot of common interest. There are several kinds of processes that determine S and R policies here. I will compare two. First, biological evolution establishes the basic architecture responsible for memory. Second, there is also a role for deliberate choice based on an agent's own preferences, which need not reflect their evolutionary "interests."

In the operation of biological evolution here, we expect a lot of common interest. It is hard to come up with a case where the biological interests of different stages of the same organism might diverge, though it may not be impossible. Given that the utility profile of an agent is affected by evolution of the same kind, we expect a fair amount of agreement over stages with respect to the preferences of the whole agent, too.[7] But there is also the possibility of divergence. One source of this is the simple fact that a psychological profile sufficient to determine preferences and choices of the relevant kind exists locally to a time. This sets up the possibility of a clash of interests across temporal stages. You might believe now that your later self will prefer outcomes that you do not want pursued. In response, you might "tie yourself to the mast," like Ulysses, or instead try to control the information flow to your future self. This is the raw material for interesting science fiction stories. (A new one, a remake of the 1990 film *Total Recall*, is coming out later this year.) In other versions of this talk I spend some time on this topic, but today I will leave it to imagination and move on.[8]

---

[7] See Sterelny's "From Fitness to Utility" on the convergences and divergences between those two.

[8] Robert Trivers' recent work on the idea that our minds are rife with socially adaptive self-deception is relevant here also.

3.2. Separation

I said earlier that a clear case of the SRC has clear separations between sender or producer, sign, and reader or user. In thinking about how the mind works, a view of thought as involving *internal representation* has long been attractive. But does a mental representation require a mental reader? If so, the idea of mental representation looks pseudo-explanatory, perhaps generating a regress, because we are assuming smartness and comprehension in the reader mechanism.

These questions have been especially sharp in the history of thinking about memory. Plato discussed a "wax tablet" model of memory in the *Theaetetus*, and according to Carruthers and Danziger, who have worked on the history, inscription has been the "master metaphor" for thinking about memory in the Western tradition ever since.[9] Associationism, from at least the time of Hume, sought a more obviously naturalistic, quasi-physical, model. This project was taken over by neurobiology. A feature of much mainstream neuroscience has been a rejection of more literal applications of the inscription model, with its implied separation between representation and reader. Christof Koch, in 1999, summarized this view by saying that in the brain "memory is everywhere, intermixed with computational elements" (p. 471). The brain is plastic – experience *affects* it, especially with respect to synaptic connections, and this has consequences for how later experience is handled. But neuroscience often avoids a model in which memory is stored and then read by a distinct reader device.

Against this background, Randy Gallistel has recently argued (especially in a 2010 book, *Memory and the Computational Brain*, co-authored with Adam King), that the brain must contain a "read-write" memory, even though no neural basis for it has yet been found. In my terms, a read-write process is a sender-receiver process, laid out in time. Gallistel and King's argument is that various behaviors that many animals routinely perform, especially behaviors involving navigation, have known computational requirements, and these include a read-write memory, a memory of roughly the kind seen in an ordinary computer. Essentially, the old inscription model must be basically right, and closer to the truth than neurobiology textbooks today.

This argument has foundational importance for the philosophy of mind. Most philosophers have got used to the idea that the rise of computer technology in the mid 20th century answered or dissolved regress arguments against a

[9] See Carruthers' *The Book of Memory* (1992), and Danziger's *Marking the Mind* (2008).

representational view of the mind. But *how* did computers do this? A standard view, expressed clearly at the crucial time by Dennett (in a 1977 review of Fodor's *Language of Thought*), is that computers showed us that readers are not needed. Computer technology shows that there can be representation without the familiar sign-reader relationships that generate regresses. Another view, drawing on Gallistel, is that computers do indeed solve the regress problem, but not in the way people think. A computer has a sign/reader distinction in its hardware, in the division between memory and processor. Computers solved the regress problem not by showing that representations don't need readers, but by showing that representation/reader configurations can be mechanized, and can be very powerful once mechanized.

In a review of Gallistel and King's book, John Donohoe expresses and defends the mainstream view by saying that the brain seems to use a "write-only" form of memory. Gallistel replied that if there was really a device like that, it would be useless: "if a machine truly cannot read what it has written, then there is no point in its having written it, because what has been written cannot influence its behavior in any way."[10] One can see a kind of cross-talk going on here, with Gallistel talking inside the S-R model and Donohoe talking outside it. The idea that Donohoe was reaching for might be better described as a *write-activate* memory; the mainstream view holds that special processes in the brain change its structure as a consequence of experience (especially through "long-term potentiation" of synapses), but these changes can affect later behavior without being read. Dedicated mechanisms install the memory, but no reader is needed to get it out.

I don't know which side of this debate is right. In my terms, this is a debate about whether memory in the brain involves a clear sender-receiver structure across the temporal dimension, or a marginal one.

I will briefly make another connection, to the philosophy of biology. It has long been appealing to describe genes and genomes as "carrying information," perhaps as "representing" things, but it is hard to tell whether this is a well-motivated description or just a vague metaphor that may contain illusions of explanation. One way to approach this is to apply something like a sender-receiver model to genes. This has been done in different ways by Nick Shea (2007) and by Carl Bergstrom and Martin Rosvall (2010).[11] I will not discuss the details of these

[10] The exchange is in *Behavior and Philosophy*, 2010.
[11] Shea's "Representation in the Genome...", Bergstrom and Rosvall's "The Transmission Sense of Information," both in *Biology and Philosophy*.

proposals here, but they tend to run into difficulties on one side or the other – the sender side or the receiver side. Looking within the cell, the "reading" of genes is a fairly well-defined matter. Cells do treat genes as information repositories. But who is the sender? We might instead step back from the cell level and see the parents of a whole organism as senders of a message, but then, as in Shea, we end up with a dubious receiver. Shea's receiver is the "developmental system" in the offspring, which I think is too much of an abstraction to be a good reader or receiver. This seems to be a case where a natural system has *some* match to the SR model, but it is a mistake to force it into those terms too literally. Marginal cases are one natural product, something to understand in their own right. Perhaps genetic systems are the flipside of the brain in one respect – or at least, they are the flipside according to the mainstream view rejected by Gallistel. Whereas brains have a *write-activate* memory, genetic systems are a case of *evolve-read* memory, without a clear sender or writer. It is interesting to think about when evolution might tend to build one of these rather than the other, and rather than a paradigm SR system.

3.3. Flexibility, inference, and reconstruction

Memory is both a personal and subpersonal matter. Gallistel's arguments are concerned with subpersonal processes, deep in the machinery. Lastly I will look at some ideas that apply at both levels.

The simplest sender-receiver system, from Lewis, is one where S and R apply fixed rules, yielding a definite function from states to acts. Lewis said the resulting signs are "neutral" in content, both indicative and imperative. Some personal-level memory processes make use of natural language to express ideas. So some remembered contents have semantic properties with the definiteness of natural language propositions, and are indicative or imperative. A functional perspective can be taken on this distinction, and it can be explored first in the case of notes-to-self and other external marks. Sometimes in external reminders we use indicatives, sometimes imperatives: *meeting is at noon* versus *pick up the car*. In choosing between them, the informational asymmetries discussed earlier are relevant. If you know your later self will have additional information about the context of action, that motivates use of indicatives.

These design arguments apply also to internal memory, including internal memory traces whose content is not derivative on public language sentences. When a later self (receiver) will have more information, it is better to allow receiver

discretion, and this goes with – either as reductive explanation or not – the use of indicative signs. That is the first "transition" in the sequence discussed earlier in this paper, the transition from an inflexible sender and receiver with "neutral" contents to indicative signs and receiver discretion.

As in the between-agent case, a receiver's flexibility need not be restricted to working out how to act, though. It can also figure in their interpretation of the marks made available by the sender. A smart receiver can make use of incomplete and non-conventional signs. This creates opportunities for the sender. They can economize and improvise, modifying standardized signs to take advantage of the receiver's smartness. The sender can be telegraphic, and the receiver can be creative.

Here is another simple case involving external memory: I read books in the sun and make notes with a laptop whose screen cannot compete with the sun. So I can't see what I'm typing. Every third letter is a typo, but it is not hard to use the script to reconstruct later what I wanted to write down. There is no invention of abbreviations, just a loss of detail within a standard linguistic code.[12] The marks are effective because they are a prompt to *inference*, and evidence that can be used to constrain the inference. We have reached the phenomena that motivate what Sperber and Wilson call the "inferential model" as opposed to the "code model" of communication.[13] This is the second transition discussed earlier – from discretionary use of standardized signs to improvised signs that rely on receiver inference.

Let's now apply these "design" arguments to internal phenomena, where introspection is often unreliable.

Ulrich Neisser, a cognitive psychologist, introduced a famous analogy in 1967. Recalling facts from memory is like what a paleontologist does when they reconstruct a dinosaur from a few bones. The bones are not the dinosaur, or a representation of one, but they can be used as evidence from which to reconstruct the whole. They are

---

[12] More extreme cases are seen with practices like leaving objects in odd locations as reminders that *something* has to be done. Here the sign has no relation at all to the relevant acts or states, and its function is just to get the receiver to think: I am supposed to remember to do *something*.

[13] Sperber and Wilson: "Mary and Peter are sitting on a park bench. He leans back, which alters her view. By leaning back, he modifies her cognitive environment; he reveals to her certain phenomena, which she may look at or not, and describe to herself in different ways. Why should she pay attention to one phenomenon rather than another, or describe it to herself in one way rather than another? ...

Imagine, for instance, that as a result of Peter's leaning back she can see, among other things, three people: an ice-cream vendor who she had noticed before when she sat down on the bench, an ordinary stroller who she has never seen before, and her acquaintance William, who is coming towards them and is a dreadful bore."

traces that a smart interpreter can make use of. "Out of a few stored bone chips, we remember a dinosaur." Contrasts can be made with memory as mummification, or freezing; memory is not a matter of storing and preserving a whole, then retrieving it and bringing it back to life, but instead reconstructing it from traces or scraps. This approach is now mainstream in cognitive psychology. (A relevant review is Daniel Schachter's *In Search of Memory*, 1996). Experiments within this approach often look at the causal role of the present-day context on what is remembered, such as the "probe" used. They also look at the role of recall events that have happened *between* the present time and the target event, and events between the present time and the target event that have affected your relationship to objects and people involved. *Reconstruction rather than reproduction* is a catchphrase for memory work.

A early hero of this tradition is Frederic Bartlett, writing in the 1930s.

> The first notion to get rid of is that memory is primarily or literally reduplicative, or reproductive. In a world of constantly changing environment, literal recall is extraordinarily unimportant. It is with remembering as it is with the stroke in a skilled game. We may fancy that we are repeating a series of movements learned a long time before from a text-book or from a teacher. But motion study shows that in fact we build up the stroke afresh on a basis of the immediately preceding balance of postures and the momentary needs of the game. Every time we make it, it has its own characteristics.
>
> Frederic Bartlett, *Remembering*, 1932, Ch 6

Compare this to the Ludlow quote about language I gave earlier.

> [H]uman languages are one-off things that we build “on the fly” on a conversation-by-conversation basis; we can call these one-off fleeting languages *microlanguages*. Importantly, this picture rejects the idea that words are relatively stable things with fixed meanings that we come to learn. Rather, word meanings themselves are dynamic — they shift from microlanguage to microlanguage.
>
> Peter Ludlow, "The Living Word," NYT April 23, 2012.

How often are these reconstructions factually accurate? This brings us to the most contentious areas of memory research – the role of memory in psychoanalysis, recovered memory in abuse cases, and testimony research. Daniel Schachter, in the review I cited earlier, says that autobiographical memory is mostly fairly accurate, in coarse-grain though not in fine details, within fairly normal circumstances of recall.

In unusual recall circumstances – such as hypnosis, and coaching during therapy and psychoanalysis – just about anything can come out.

Let's apply the functional perspective outlined earlier. In personal-level memory phenomena there is a smart receiver, who is able to make inferences and fill in gaps. A sender can make use of these facts; the sender can economize, laying down less than the whole dinosaur. This is not merely a situation where memory can *fade* to a few scraps of bone, but all the earlier self *needs* to lay down are bones. We can *expect* this economizing as a sender behavior where whole-organism intelligence will be brought to bear on the receiver side.

Sometimes a later self will add what looks like too much content. Outside of cases involving unusual recall circumstances, this often seems to involve the shaping of events into a more coherent narrative. In memory and in other contexts, people are attracted to stories that make sense, especially with respect to the arc of their lives and how things fit together. That tendency seems to operate strongly when dealing with memory traces.

This over-stepping and over-shaping, in turn, has at least two possible explanations, within the framework used here.[14]

(i) Distortion of memory as a maladaptive byproduct. Creative distortion is the cost paid for the economies and efficiencies that come from the third mode of sender-receiver interaction, the mode where the receiver is free to interpret traces and the sender is designed to make use of that fact. An intrusion of the narrative-seeking urge into processes of recall leads to unwanted side-effects.

(ii) Adaptive explanation. Accurate signaling between stages is one thing that organisms benefit from, but there are other cognitive benefits of memory. We benefit from imposing a coherent story on our lives, even when this compromises factual accuracy.[15] The traces left by the sender are designed for this use, as well as for accurate reconstruction.

There are several other angles that could be followed up here. The SR model gives a way of thinking about memory that moves away from the simple idea of

[14] There may also be the possibility of applying a conflict of interest model here, but it is not necessary to the options below.

[15] Schachter's discussions of amnesia in *In Search of Memory* emphasize the extreme cost of a loss of a sense of a self*,* with projects to pursue, that comes from loss of memory.

"storage," and places departures from that idea into an adaptive context. In this talk I have not looked in detail at how the content of signs is determined in SR systems, and at how different grades of discretion and receiver rationality affect those matters. These relationships take interesting forms in the case of temporally organized SR systems such as memory.