CHAPTER EIGHT

Representational Genera

A *genus* of representation is a general kind, within which there can be more specific kinds, importantly different from one another, yet generically alike. The level of generality intended can be indicated by example. Natural languages, logical calculi, and computer programming languages, as well as numerous more specialized notations, are all interestingly different species; but they are generically alike in being broadly *language-like* or *logical* in character. By contrast, pictures, though equally representational, are not linguistic at all, even in this broad sense; rather, they, along with maps, scale models, analog computers, and at least some graphs, charts, and diagrams, are species in another genus of broadly *image-like* or *iconic* representations. So the level of generality intended for representational genera is that of logical versus iconic representations, thus broadly construed.

The motive for the notion lies in the possibility of raising two kinds of question. First: on what basis are such genera gathered and distinguished? That is, what sort of likeness or disparity between two representational species determines whether they are in the same or different genera? Given an account of this basis, it should be possible to delineate the distinctive "essence" of each genus. Second: besides the two familiar genera, are there any others? In particular, is so-called *distributed* representation a separate genus, on a par with logical and iconic representation, yet as different from each as they are from one another? And, if so, what is its generic essence?

Inasmuch as distributed representations are still relatively strange and (hence?) unintuitive, the latter questions can best be approached by way of the former. Accordingly, even an investigation into the essence of distributed representation must be guided in large part by intuitions grounded in the better-known logical and iconic genera—

in effect, bootstrapping these up to a common level, which can then be extended and applied to less familiar cases. The considerations that follow are even more than usually tentative and exploratory; the results are at best preliminary and incomplete, perhaps much worse.

1 Representation: the "family" of the genera

An explicit account of representation as such will not be necessary; that is, we can get along without a prior definition of the "family" within which the genera are to be distinguished. A few sketchy and dogmatic remarks, however, may provide some useful orientation, as well as places to hang some terminological stipulations.

A sophisticated system (organism) designed (evolved) to maximize some end (such as survival) must in general adjust its behavior to specific features, structures, or configurations of its environment in ways that could not have been fully prearranged in its design. If the relevant features are reliably present and manifest to the system (via some signal) whenever the adjustments must be made, then they need not be represented. Thus, plants that track the sun with their leaves needn't represent it or its position, because the tracking can be guided directly by the sun itself. But if the relevant features are not always present (manifest), then they can, at least in some cases, be represented; that is, something else can stand in for them, with the power to guide behavior in their stead. That which stands in for something else in this way is a *representation*; that which it stands in for is its *content*;[1] and its standing in for that content is *representing* it.

As so far described, "standing in for" could be quite inflexible and ad hoc; for instance, triggered gastric juices might keep a primitive predator on the prowl, even when it momentarily loses a scent—thus standing in for the scent. Here, however, we will reserve the term 'representation' for those stand-ins that function in virtue of a general *representational scheme* such that: (i) a variety of possible contents can be represented by a corresponding variety of possible representations; (ii) what any given representation (item, pattern, state, event, ...) represents is determined in some consistent or systematic way by the scheme;[2] and (iii) there are proper (and improper) ways of producing, maintaining, modifying, and/or using the various representations under various environmental and other conditions. (This characterization is intended to be neutral not only among genera, but also between internal and external representations, and between natural and artificial schemes.)

Since the content of a given representation is determined by its scheme (and since the point of the facility is to be able to represent what isn't present or currently accessible), it is possible for representations to misrepresent. What this amounts to will vary with the specific scheme, and even more with its genus; but it must hark back eventually to the possibility of the system(s) using it being *misguided* in their attempted adjustments to the features of the world. But misrepresentation should not be confused with improper deployment on the part of the using system, nor bad luck in the results. These can diverge in virtue of the fundamental holism underlying what can count as a representation at all: the scheme must be such that, properly produced and used, its representations will, under normal conditions, guide the system successfully, on the whole. In case conditions are, in one way or another, not normal, however, then a representing system can misrepresent without in any way malfunctioning.

2 Canonical accounts of the genera

Analyzing representation in terms of a relational structure—that which represents–representing–that which is represented—suggests looking for the distinctive essences of the genera in one or another of those elements. But they clearly do not lie in the representative tokens alone. One cannot simply say, for instance, that linguistic tokens are essentially digital, whereas images are analog, since, in the first place, speech is to some extent analog, and in the second, images are often handled digitally. More to the point, however, paradigm representations of *any* genus—logical, iconic, distributed, or whatever—can be recorded without loss[3] in strings of bits; and certainly no properties of bit strings as such can differentiate in general between the sorts of representation they record. And, on the face of it, there seems equally little hope of differentiating the genera on the basis of their represented contents. Thus, a police officer's report might well include both descriptions and photographs of one and the same crime scene; and, *prima facie*, there could be distributed representations of it as well.

Apparently, then, the essential differentia must be sought in the nature of the representing itself, the relation between representations and their contents; and, indeed, standard characterizations of all three genera take this form. I, on the contrary, however, will argue that the distinctions cannot be made out in terms of the representing relation (but are to be found in what is represented—the contents—after all). In order to appreciate what's wrong with the standard approach, we will need first to outline how it is supposed to work for each genus. These outlines can be called the *canonical accounts,* because they are what almost everybody expects almost everybody else to believe. It will not be necessary to formulate them precisely, since the underlying problem is common and structural.

Logical representations are distinctive in that they have a (generative) *compositional semantics*. This means that complete logical tokens (sentences, well-formed formulae, production rules, and so on) are complex structures, each with a recursively specifiable syntax and determinate atomic constituents, such that the semantic significance of the whole is determined by its syntax and the semantic significances of its constituents (perhaps with some situational parameters). 'Semantic significance' is here used as an umbrella term, meant to include, without being limited to, the represented contents (what the tokens represent). The respective contributions of the possible structures and constituents are fixed arbitrarily by the scheme (for instance, a language); but, given these (and any relevant parameters), the significance of the compound is not at all arbitrary.

Iconic representations are distinctive in representing their contents by virtue of being somehow *isomorphic* to them. In many familiar cases, like pictures and scale models, the isomorphism is obvious enough to strike the observer as *resemblance*. But that's not required: the isomorphisms that determine the representational contents of graphs, wiring diagrams, and analog computers are often so abstract or mathematical that we wouldn't naturally call them resemblances (though, of course, in a sense, they are). Note that there are many different kinds of isomorphism, and the ones that are relevant to any particular representation are all and only those determined by the scheme to which it belongs. Thus a picture token that happens to be entirely monochromatic may or may not represent its object as likewise monochromatic, depending on whether it belongs to a monochromatic or full-color scheme. Which isomorphisms a given scheme employs is initially arbitrary or conventional; but once they are fixed, the contents of particular tokens are not arbitrary.

Distributed representations are distinctive in that each portion of the token participates (in some broadly egalitarian way) in the representing of each portion of the represented contents. For this to make sense, there must be antecedent notions of portions, both of the tokens and of the contents; and then the idea is that the representings of the content portions are all spread out over the token portions and superimposed on one another—whence the term "distributed". What these respective portions are, and how various possible content portions can be (simultaneously) determined by various possible token portions, is, of course, settled by the particular scheme (which is arbitrary); but, given a scheme, it is not at all arbitrary what any particular token represents.[4]

Each canonical account focuses on a distinctive sort of relation between representing tokens and their represented contents—namely, a relation that is systematically determined by complex structures and a primitive vocabulary, a relation that is based on having a common abstract form, and a relation that somehow has its components all mixed up (spread out and then superimposed). There is, of course, no suggestion that these possibilities are exhaustive; it is an interesting and important question whether there are or could be other genera of representation, and if so what they are. As canonical accounts of essences, however, they should at least be mutually exclusive, so as to distinguish their respective genera. Unfortunately, they are not.

3 Problems with the canonical accounts

This superficial difficulty with the canonical accounts will be exploited, via a three-stage argument, as a symptom of an underlying common failing. The first stage is a group of three counter-example sketches, designed to suggest that each account admits cases that properly belong to another genus, and hence that none of them successfully captures what is distinctive of its own genus. As with the accounts themselves, the counter-examples will be presented informally, since the point is not really to get any of them right, but rather to "see through" them, in preparation for a completely different tack. The second stage will further that cause by advancing a trio of more outlandish counter-theses, which are therefore also more transparent in their operation. In particular, it should be obvious that they all trade on failing to respect one and the same fundamental distinction. The final stage then re-examines the counter-examples from stage one in the light of this distinction, and recruits them to show that the failures at stage two lie not with the outlandish counter-theses, but with the canonical accounts themselves.

So, stage one begins with three counter-examples. Consider first a designer's floor plan, a kind of two-dimensional scale model of a room, with movable cardboard cutouts for the various furnishings, fixtures, and pathways. There are explicit rules for composing these primitives into well-formed rooms: the pieces must lie flat, they shouldn't overlap, pathways have to be connected, and the like; and, according to essentially the same rules, these room plans can then be assembled into building plans, which in turn can make up city plans, and so on (recursively), with no principled limit. For a digital variant, arbitrarily define a grid (set of axes and unit pixels) on the Cartesian plane, and assign distinct colors to a finite set of landscape features (blue to water, green to forest, and so on); then maps can be recursively defined as follows: (1) a grid with a single colored pixel is a map; and (2) a consolidation of two maps (coincident axes, no pixel conflicts, all colored pixels contiguous) is a map. In each example, the represented content of the resulting plan or map is manifestly a determinate function of its compositional structure and the significances of its primitive components—just the token/content relation that was supposed to be characteristic of logical representations. Yet floor plans and maps are paradigmatically iconic.

Consider next how abstract and general the notions of form and isomorphism are, and must be, if the canonical account of the iconic genus is to accommodate such abstract representations as graphs, analog computers, and wiring diagrams. Yet, in a sufficiently abstract sense of form, it is arguable that many other (perhaps all) tokens represent their particular contents by exhibiting the same form. Wittgenstein (1921/74), for instance, once presented a theory of meaning according to which sentences represent worldly facts by "picturing" their logical structure; and, more recently, Sellars (1963/67, 1985) has proposed an account of predication based on the idea that juxtaposition of predicate symbols is essentially a flexible way of giving different properties to terms—in a manner that "maps" the properties of the referents of those terms. Regardless of whether any such approach is viable in general, if it can be made to work for even one scheme of logical representation, then the canonical account of iconic representation fails to exclude it.

Consider finally ordinary holograms: they seem to be paradigmatic examples of the spread-out-ness and superposition that are officially characteristic of distributed representation. Thus, in the original and most familiar sort of hologram, representing a scene in 3-D, each point on the holographic plate represents (in some measure) each part of the scene and each part of the scene is represented (in some measure) by each point on the plate. Yet, in a perfectly straightforward sense, such holographic representations of scenes are patently images—just as every popular article and exhibition unabashedly calls them. What's more, as is well known, holographic techniques can also be used in error-resistant encodings of logical representations.

These counter-examples are alike in suggesting that the canonical accounts are insufficiently restrictive (each admits cases that don't belong) and hence are incapable of supporting principled distinctions among the genera. One possible philosophical response would be to try ruling out the examples by somehow sharpening the accounts; another would be to give up, and concede that the alleged genera are not, after all, essentially distinct. I am arguing instead that all three canonical accounts misfire for a single underlying reason—namely, they all mislocate the generic essences in the representing relation—and, hence, the appropriate response is neither repair nor despair but fundamental reconception.

4 The representing/recording distinction

The second stage in this argument now proceeds indirectly, through a matched set of outlandish theses whose transparently spurious justifications are designed to bring out the need for a basic distinction.

Outlandish thesis #1. It is easy to "translate" any iconic or distributed representation into an equivalent logical representation—that is, a set of sentences representing the same contents. Before considering the "argument" for this thesis, we should perhaps remind ourselves why it is outlandish. Explaining perceptual recognition—how, for instance, a system can look at a scene (or picture) and produce an articulate verbal description of what it sees—is a profound outstanding problem in psychology and artificial intelligence. It has proven exceptionally difficult, and remains substantially unsolved. Yet thesis #1 implies that it should be easy. Here's how: use a simple computer-driven scanner to divide the scene (or picture) into a large number of cells, and generate on a printer, for each cell, a sentence token of the form: "The light reflected from direction (or pixel) *x*, *y* is of intensity, hue, and saturation *i*, *h*, and *s*." The supposed translations of distributed representations would be equally fatuous—for instance, a list of sentences of the form: "The weight on the connection between nodes *j* and *k* is *w*."

OUTLANDISH THESIS #2. It is easy to "translate" any logical or distributed representation into a representationally equivalent image. Like perceptual recognition, this ought to be hard—harder, for instance, than the job of a police artist, for the translator would have to draw an accurate picture of the suspect, or the scene of the crime, from the description alone, without any corrective feedback from the witness. Yet thesis #2 implies it should be easy. Here's how: Obtain a clear, written copy of the relevant description, focus a camera on it, and take a picture.[5] The comparable iconic translation of a distributed representation is only marginally less contrived: choose some sensible one-to-one correlation between the connections in a given network and the cells in a graphical array, and shade or fill each cell in proportion to the weight on the corresponding connection.[6]

OUTLANDISH THESIS #3. It is easy to "translate" any logical or iconic representation into an equivalent distributed representation. To establish thesis #3, we invoke a recent invention called *Turing's piano*. This versatile instrument has arbitrarily many notes, each a pure unmodulated sine wave, ascending in a standard chromatic scale from one hertz. Remembering, then, that any text or icon can be encoded without loss in a finite string of bits, and noting that each finite string of bits corresponds to a unique chord played on Turing's piano, we conclude that any text or icon can be encoded in such a chord. The waveform produced, however, is a perfect distributed representation, according to the canonical account: each point on the waveform (instantaneous amplitude) participates in recording all of the bits, and the recording of each bit is spread over all of the waveform. Since whatever it is that is representational in the original text or icon is somehow captured in patterns of these bits, that representing too is spread out and superimposed in the waveform.

What these irritating examples have in common is most evident in the case of a photographed inscription. While the result is, in one sense, clearly an image, it is equally clearly *not* an image of whatever the inscribed text was about; rather, it is an image of the inscription itself, but in such a way that (assuming the text is still legible in the image) the ability of the original to represent whatever it represented is preserved in the image. Obviously, however, that representing remains essentially logical in character. In other words, the logical representation has in no sense been "translated" or transformed into an iconic one; instead, it has merely been "recorded" in an iconic medium. Significantly, the recording process, unlike genuine translation, would proceed just as smoothly even if the given inscription were complete gibberish. Evidently then, we must distinguish in this case between the logical *representation* and its iconic *recording*.

The scanner-generated printout of sentences describing individual pixels is basically symmetrical, but with an important added twist. In case what is scanned is an image, then that image is clearly what the sentences describe, pixel by pixel; and, as before, assuming the descriptions are fine enough, the ability of the image to represent whatever it does is preserved, but only as iconic. Thus, the image is not translated or transformed into a logical representation, but only recorded in a logical medium.[7] (And, again, the recording itself would be equally successful even if the "image" were a contentless blob.)

The new twist emerges when we consider the companion case in which what is scanned is not an image of a scene, but the scene itself. Then, it seems, we must say that the resulting sentences describe the scene, since there is no image in play for them to describe, or to record logically. Yet, they don't describe the scene in anything like the way an articulate observer would describe it; rather, the "description" is essentially the same as what the scanner would have delivered in recording a photo of the scene. Hence, even when the sentences are generated directly from the scene, what they are doing is logically recording an *image-like* representation—a *virtual image*, we could call it—from the perspective of the scanner; and it is this which the sentences describe, pixel by pixel. This characterization is already intimated, I suspect, in calling what is represented a "scene", rather than, for instance, the facts or a state of affairs.

Turing's piano, of course, can be understood in the same way. It generates distributed *recordings* of arbitrary bit strings; indeed, multiple independent strings can be simultaneously recorded in the same chord, either interspersed or concatenated. Whether, as *representations*, these recorded strings are logical, iconic, or, perhaps, even distributed, depends on something else—it is not settled by the fact that the actual waveform records the bits and the contents (if any) in a spread out and superposed manner. What else it depends on, that is, how the representational genus is determined, remains an open question.

The common moral of the examples is now clear: we must distinguish between recording and representing, and that on two levels.

Consider once more a photo of an inscription: in favorable cases, it both represents and records that text; and, meanwhile, the text itself is a (different) representation of something else. What is the difference between recording and representing, and how is it that they are so readily confused? Fundamentally, I think, recording is a *process* of a certain sort; and to be a record is to be the result of such a process. By contrast, representing is a functional *status* or *role* of a certain sort, and to be a representation is to have that status or role. These are not the same because, in general, being the result of a recording process is neither necessary nor sufficient for having a representational status or role. On the other hand, the recording process does produce an output which is a record *of* its input; and representational status does involve one entity *standing in for* another. This formal relational parallel, plus the fact that, in certain cases (such as photography), the product of the recording process can also have the status of representing the input, therefore being simultaneously a record and a representation *of* it, is the basis of the confusion.

Recording is a generalization of copying. To count as recording, a process must be reversible in the sense that there is a complimentary process—"playing it back"—which reproduces from the record a duplicate or copy of the original. Since a copy of a copy is a copy, copying is a special case of recording (in which the record and play-back processes are the same). Recording, like copying, is only partial; only certain prespecified aspects of the original are recorded. In case what is being recorded is a representation (as a representation), then the relevant aspect, what must be reproduced on play-back, is its schematic type; the criterion for successful recording of representations is preservation of type-identity.

Recording is also essentially trivial or mechanical. Describing a scene from a picture, or drawing a picture from a description, takes experience and skill, whereas the camera or scanner "merely" records the original. This notion of triviality, however, must be delineated carefully; after all, any fool can tell a story, whereas cameras and scanners require sophisticated technology. The crucial point is that recording and playing-back are completely *witless*—the processes themselves are oblivious to content (if any) and ignorant of the world. By contrast, articulate description and artistic rendition depend crucially and thoroughly on general *background familiarity* with the represented contents—that is, on worldly experience and skill. The claim, of course, is not that representations are never produced witlessly (remember photography), but only that records always are.

An apparent exception (but one that proves the rule) is a court recorder, whose task is to produce a verbatim transcript of the oral proceedings. This task is anything but witless (so far, it can't be done by machine at all, despite substantial investment). The reason is that the phonic realization of phonemic units is highly variable, and non-disjoint; that is, the same word "sounds" quite different on different occasions, to the point of overlapping with the sounds of other words. A skilled recorder, like any speaker of the language, can identify the words correctly on the basis of knowing what "makes sense" in context—the exact opposite of witlessness. But this way of putting it shows the way to the answer. A court recorder's job is actually two-fold: identifying the spoken words (non-trivial) and then recording them (trivial). The case is essentially similar to producing a typescript from someone's handwritten scrawl, or even restoring the missing characters in a spotty photocopy: identifying the words takes wits; recording them, once identified, does not.

These considerations further suggest, though by no means establish, a sufficient condition for sameness of representational genus: if the representations of one scheme can be witlessly transformed into equivalent representations from another scheme by a general procedure, then those schemes are species of the same genus.[8] Thus, Morse and ASCII coded texts are equally logical, whereas pictorial and verbal representations differ in genus. Of course, this is not a necessary condition, since distinct natural languages cannot be witlessly intertranslated, yet they are of the same genus. Even as a sufficient condition, the proposed test is so far only hypothetically acceptable, since it depends on the plausible but unsupported assumption that cross-generic "translation" always requires wits.[9] Only later will we be able to explain and support the plausibility of this assumption, and thereby defend the suggested condition.

5 The initial problems rediagnosed

The third and final stage of the argument against the canonical accounts of the generic essences returns to the counter-examples introduced at stage one. The aim is to rediagnose these examples as showing not merely that the canonical accounts are insufficiently exclusive, but more importantly (and damningly) that they themselves confuse the representing/recording distinction.

Ordinary holographic images of scenes are the easiest case. In principle, they are no different from a Turing's piano chord produced from the output of a digitizing scanner. That is, if the hologram is prepared from an ordinary image (say, a photograph), then it is a distributed recording of that image, which, qua representation of the original scene, remains iconic—even in its recorded form. If, on the other hand, the hologram is prepared directly from the scene itself (as it must be to achieve the famous 3-D effects), then, as in the earlier treatment of the scanner, it is effectively recording a "virtual image"—which is again still an iconic representation of the scene. In other words, the distributedness in the example derives from and characterizes not the genus of the representation but rather the technology with which it is recorded.

The second counter-example, the picture theory of meaning, has the same basic structure, though in a less obvious way. Suppose we have a logical scheme of representation for which some sort of picture theory provides a plausible semantics. Then a slavish parallel to the above treatment of holographic images would go like this: the representations in the specimen scheme are indeed logical, but they are recorded in an iconic form. What's more, since there have been no actual prior sentences to record iconically, we must say that the recordings are of "virtual sentences"—something like the *facts* pictured. Now, surprisingly perhaps, I think this account is basically correct. But it does call for some explanation.

Recall the strategy of the original example: first it is pointed out that isomorphism is a very general notion, and indeed must be so if the account is to capture all kinds of icons; and then it is observed that such a general notion might apply equally well to the relation between sentences and facts. But this latter is just the tip of the iceberg. Isomorphisms are everywhere: a gunslinger's notches and his trail of victims, magnetic bit patterns and character strings, cash register tapes and the day's transactions, chess transcripts and the game's moves, and so on. All of these are familiar forms of record keeping, and none of them is plausibly an image. Thus, isomorphism is a common basis for making records, and not a distinctive mark of iconic representation. (A reversible witless process is a likely candidate for a function defining an isomorphism.) Hence, a general isomorphism between the tokens of some scheme and their contents would not alone make that scheme iconic; for it might instead just be the way the contents, understood as "virtual tokens", are recorded.

But surely this is strained: we have actual logical representations (sentences) allegedly both representing and isomorphically picturing "virtual logical representations" (the "facts"); yet, the isomorphism is not supposed to have anything to do with the representing. Rather, the sentences are said also to record those same facts; and the isomorphism is supposed to be relevant only to this recording. But what then is the sense of calling the facts "virtual *logical* representations"? And what witless process could produce such a recording; indeed, isn't describing a situation a paradigm of a process that is not witless?

The appearance of paradox can be alleviated by reexamining two earlier examples. Remember first the task of the court recorder, which is far from witless as a whole, but which can be resolved into two subtasks: identifying the oral tokens, and then recording them. Once it is appreciated that only the former requires background familiarity with the contents and the world, then the recording itself can be seen as witless after all. Then return to the case of the scanner producing pixel-by-pixel descriptions, either from a photo of a scene or directly from the scene itself. It was comparison of these that suggested the notion of recording a virtual token in the first place; and, conveniently, a scanner working directly from a scene was still a thoroughly witless process. But that process too can be conceived as a composite of two others, though in this case both witless: producing a photo-like image of a scene, and recording that image pixel by pixel.

Consolidating the two examples allows us to regard describing a situation (in a language for which a picture theory is workable) as also a composite process: identifying the pictured facts, and then verbally recording them. As with the scanner, the first step yields a virtual token, but, as with the court recorder, that step is not witless—hence neither is the composite. But the second step, the recording itself, remains witless in every case. Needless to say, this story makes only as much sense as the notion of facts as virtual sentences; but that much is a burden of the picture theory already.

Finally, the example of recursively generated floor plans and maps is at least as complicated and interesting as the preceding. Note first that the generative composition is, in each case, a *process*, starting from given primitives, proceeding by formally specifiable, rule-governed steps, and yielding the representation as a product. It is precisely this process, moreover, that is cited when confronting these icons with the canonical account of logical representations. To be sure, the constructive process is not at all like that of recording a representation, whether actual or virtual; it is, however, quite like the reverse process of recovering a representation from a record of it.

To see this, consider the intermediate case of an architectural drawing created by a computer-aided design system: it is stored within the system as a set of line specifications, or, equivalently, line-drawing commands, which can be sent to a pen-plotter or display driver to execute. Now, those internal specifications are just as much a logical record of an iconic representation as are our earlier printed sentences describing individual pixels; and commanding the plotter to draw the lines is just the (witless) process of "recovering" that image. The fact that the image may never have existed in visible form before does not mean that the process of drawing it is not the inverse of recording. The commands to the plotter, however, are exactly analogous to the formal specifications of the steps by which the floor plan or map may be constructed—except that the step specifications need not actually be written down, let alone all together in advance.

That suggests the following diagnosis: recursively generated maps or floor plans are not therefore logical representations, but rather are icons witlessly recovered from (possibly "virtual") logical recordings of them. To be sure, the idea of playing back virtual recordings has a certain air of hocus pocus; but that doesn't affect the point. We want to know what it is about recursively specified maps that gives them a semblance of being logical or language-like, even though they're not; and the answer is that the recursive specification itself functions in the account just as if it were a logical recording from which the image is being restored.

Each canonical account characterizes its genus in terms of a distinctive sort of representing relation between the representations and what they represent—roughly compositional semantics, structural isomorphism, and spread-out superposition. Our diagnoses of the three counter-examples, however, have revealed these distinctive relations to be indicative more of certain recording devices or processes than of the respective representational genera. Of course, one might still undertake to revise the definitions and/or rebut the diagnoses; but the consistent pattern of results invites a more radical response—namely, to abandon the canonical accounts completely, and seek the distinctive marks of the genera elsewhere than in the representing relation. Since the argument that a bit string can be a (recorded) representation in any genus decisively rules out finding the marks in the representations themselves, I now want to explore further the prospects of finding them in what is represented—the contents of the representations.

6 Skeletal versus fleshed-out content

The first question to ask is: What, strictly speaking, *are* the contents of various representations? Exactly what objects, features, or configurations in the world are represented? Often, in real-life communication, only a little need be indicated, in order to apprise the recipient of much about the world. In practice, a little representation goes a long way. But that's because the recipient is already in a position to learn much from just a little new information—perhaps by already "knowing" a lot about the current situation and/or about the world in general. How much of what derives from this "background familiarity" belongs in the represented contents as such?

In one sense, all or most of what a recipient could be expected to learn from a communicated representation could be deemed "contained" therein—that is, as part of the message. Unfortunately, such a liberal attitude is problematic in the context of qualitatively differentiating the contents characteristic of the respective genera. For, if the recipient's broader understanding is mediated by other (presumably internal[10]) representations, themselves belonging to a genus different from that of the representation being examined, then the contents broadly construed would confusedly exhibit characteristics of both genera. Since it is not possible in general to know about or control for the effects of all such mediating representations, there is no hope of finding the essences of the genera in the character of their contents, unless these contents can be delimited more narrowly.

By way of anticipation, then, let's call the strict content of a representation, that not augmented or mediated by any other, its *skeletal* or *bare-bones* content. The metaphor is intended to suggest that the full-blooded contents of everyday representations are shaped and supported by their skeletal contents, but that they are (or can be) fleshed out and enlivened through other influences. How can this distinction be made out?

Consider again the police officer's report from the scene of a crime—including both descriptions and photographs. Clearly, the room was searched; clearly, there was a violent struggle, resulting in the death of the victim; clearly, the victim was robbed; and much else besides, as anyone can tell. But, reading more closely, the descriptive report does not actually say that the room was searched; it says only that all the drawers were opened, with the clothing removed and strewn about. It says there were signs of a struggle, including spilled drinks, broken furniture, and of course, a dead body; but it nowhere characterizes that struggle as violent. Finally, though it does point out that the decedent's wallet was found on the floor, with all the cash and credit cards missing, the report omits any mention of robbery.

Of course, the officer has made numerous assumptions in describing the scattered clothing as the "removed from the drawers", in saying the cash and cards are "missing", and so on; but that's irrelevant here. What matters, in delimiting the skeletal contents of the descriptions, are rather the natural "assumptions" of the *reader*. Thus, however natural it is, perhaps even part of what the speaker intended, and prerequisite to understanding the text, for the reader to appreciate that the drawers were searched, the struggle violent, and the missing money stolen, the fact remains that none of these is strictly contained in the report itself—any more than the lurid and sensational surmises that will make of it a journalistic bonanza. The content of what is strictly said—said "in so many words"—is intuitively what we mean by *skeletal* content.

There are many difficulties with this distinction; of these, I will discuss three briefly (none adequately). First, the skeletal/fleshed-out distinction should be located relative to other and better-known distinctions, such as literal versus figurative, and explicit versus implicit. Clearly it does not coincide with the former (even adjusting for the fact that figurative excludes literal, whereas fleshed-out includes skeletal). Thus, none of the reader's natural assumptions (that the missing money was stolen, and so on) are in any sense figurative; yet they are not part of the bare-bones content. It might be, however, that (for logical representations) a figure must always be shaped and supported by a skeleton, hence that the skeletal is somehow a subset of the literal.

The explicit/implicit distinction is closer, in that (again, for logical representations) the skeletal may roughly coincide with the explicit; but whether the rest of the full-blooded content can be regarded as implicit depends on how carefully that word is used. It is not irrelevant that the connotations of 'implicit' ("folded within") run exactly counter to those of 'fleshed out': the one suggests "already there but hidden", the other "added on". More specifically, the implicit is what is implied, in the sense that it can be got or brought out by inference. But, even if this is taken broadly to include material implication (such that, from a formal point of view, common knowledge would be functioning as a suppressed premise), the merely implicit falls far short of full-blooded content. Only if "implication" is stretched to the full gamut of what can be expected from conversational skills, topical associations, critical judgements, affective responses, and so on, can it hope to encompass the sort of content routinely conveyed in the newspaper, in the board room, and over the back fence. But, quite apart from whether (as I think) that's an overstretching of the word, it remains the case that implication, like trope, is proper to logical representation, whereas the bare-bones/fleshed-out distinction is intended to apply in any genus.

Yet another potentially confusing distinction in the vicinity might be likened to that between a pig and its rider: some representations incorporate or "carry on their shoulders" others that work according to different schemes, maybe even in different genera. Here I do not have in mind cases where, within one representation, another is represented (the protesters' banners visible in the photos of the rally), or encoded (the secret formula hidden in an innocent love letter), but rather "piggyback" devices that systematically belong and contribute to the representations that carry them. Perhaps the clearest examples are the legends and descriptive tokens that often appear on maps; but the standardized symbols in medieval paintings, and the labels in political cartoons, are essentially comparable. Onomatopoeia and "graphic" textual effects (advertising copy that not only touts speed, but itself "looks" fast) are presumably analogous, in the other direction. It seems to me that the additional content in piggyback devices might in principle itself sustain a skeletal/fleshed-out distinction, and that none of it should be counted in the content (even the full-blooded content) of the underlying representation. What we have instead is a sort of hybrid, with separate but related content on two distinct levels.

Second, the idea of "fleshing out" skeletal content suggests interlocutors who, over and above grasping the content of a representation as dictated by its scheme, also bring into play general background knowledge and familiar at-home-ness in the world at large. Put this way, however, the notion seems to presuppose a scheme/background distinction, which may in turn seem tantamount to the sort of language/theory distinction that Quine warned us not to draw. But the worry is misguided. What Quine argued against is the possibility of a systematic distinction, in the grounds for what we say, between arbitrary choice or convention and empirical evidence or theory; in particular, the language in which a theory is expressed cannot be separated from that theory in such a way that some theses are defensible on linguistic grounds alone, with no admixture of theory. But that is not at all incompatible with distinguishing between what a particular thesis says "in so many words" (according to the language-cum-theory of which it is a part), and what its concrete "implications" are in the current circumstances (according to that same language-cum-theory). Still less, therefore is it incompatible with the distinction between skeletal and full-blooded content, inasmuch as the latter extends even beyond implications.

Third, and most serious: artificial intelligence research has established that natural language competence is impossible without a common-sense grasp of real-live flesh-and-blood content; that is, systems that lack such a grasp fail to understand natural discourse at all. But if language with skeletal content alone is unintelligible, then perhaps skeletal content, as a kind of abstraction from "living language", makes no sense. This argument effectively assumes, however, that each representational scheme must make sense on its own, apart from all others; but, once exposed, that assumption seems to have little ground to stand on. Why, for instance, couldn't one "living" scheme be *parasitic* on another, or two schemes be *symbiotic*? Thus, it might be granted that natural language is viable (as a means of communication, or whatever) only with its full-blooded everyday content, but this be attributed to its dependence on (or interdependence with) some other scheme of representation—presumably internal to the language users. In other words, the full content of a discourse, in terms of which it is workable at all, is simultaneously a function of two determining factors: the skeletal content of those linguistic tokens themselves, plus whatever else the relevant sensible speakers of that language can count on one another to grasp in that context. That the latter is essential in practice does not show that the former is impossible in theory, or indeed inessential.

The suggestion that language depends on inner representations has, of course, a long history. But traditionally, the idea has been that thought is the original locus of all contentfulness, and that linguistic tokens acquire contents only by having thought contents somehow conferred on them. The present notion of "symbiosis", however, is much less definite. In the first place, skeletal linguistic contentfulness need not be dependent on internal representations at all; and even fleshed-out content is dependent only in the sense that it arises in the linguistic practices of interlocutors whose usage relies also on other representations. These do not so much as imply that linguistic contents are of the same sort as those of internal representations, let alone that the former are merely conferred duplicates of the latter.

Before attempting to characterize the skeletal contents of logical representations more specifically, let's ask in a general way whether a comparable distinction can be made out for iconic representations. What, for instance, would be the bare-bones contents of the police officer's photos? On the one hand, the pictures reveal much that is omitted from the descriptions, like the shape of the bloodstain on the carpet, the color of the victim's shirt, and the cockroach in the corner. On the other hand, much that the officer explicitly said is not represented in the photographs. For example, not only do they not show the money and credit cards as stolen, they don't even show them as missing. More interestingly, the pictures don't actually represent the victim as dead, or, for that matter, as a victim—though, of course, we would be likely to appreciate these facts, if we saw the pictures.

How far does this go? The same reasoning, pressed to its limit, implies that, bare bones, the photographs don't even represent a human body lying on the floor in a pool of blood, surrounded by scattered clothing and broken furniture. These are features that an ordinary person, relying on common background familiarity with the world, could easily "tell" about the situation depicted. But, as the perceptual recognition research cited earlier has demonstrated, no system innocent of the world, lacking common sense, is capable of such telling. The ultimate conclusion would then be that all the photos "strictly" represent is certain variations of incident light with respect to direction—taking seriously, in effect, the pun in photo*graph*. Of course, the pictures surely do represent the scene of a violent robbery, just as much as the verbal reports describe one. The point is not to deny the obvious, but rather to distinguish, within the undeniable contents of everyday representations, a substructure that is skeletal—in the special sense that it does not draw upon the user's antecedent familiarity with the situation and the world, and hence cannot exhibit any admixture of characteristics from other (possibly symbiotic) representational schemes.

7 Distinguishing genera by contents

The motive for distinguishing skeletal from full-blooded content is to clear the way for characterizing the essences of the genera in terms of the natures of their respective contents. Fleshed-out content is not a candidate for this role, inasmuch as it may confound the traits of more than one scheme. Hence, it could at best be the general characters of the respective skeletal contents that are distinctive of the representational genera.

What is the distinctive structure of the skeletal contents of logical representations? The mainstream tradition in logic has always taken atomic sentences to be about the properties of (or relations among) things. Modern formal semantics generally remains within this tradition by explicating property instantiation in terms of set membership and/or function application, and therewith mathematically extending the approach not only to quantification and sentential composition, but also (so it is hoped) to tense, aspect, modality, adverbial modification, propositional attitude ascription, and what have you. Somewhat further afield, but still logical, are calculi of parts and wholes, systems representing condition/action pairs, notations for allocations of resources to tasks, and so on. Of course, these are not incompatible: conditions can be objects having properties, tasks can have others as parts—indeed, in its way, natural language can handle them all.

It is awfully tempting to say that these contents of atomic sentences are (possible$^{\text{II}}$) atomic facts, somehow composed of, for instance, objects and properties; and that the contents of molecular sentences are molecular facts, composed of, or in some other way determined by, atomic facts. The difficulties in spelling out what sort of structures or features of the world such facts might be, however, are well-known and formidable. Thus, even if it is intelligible to say that a certain object's having a certain property is a structure or feature of the world, it remains unclear what kind of structure or feature it would be for an object *not* to have a property, or for it to have *either* one property *or* another, or for it to be the case that *at least one* object has some property. To be sure, each such "complex fact" is settled in any given instance by some or all of the atomic facts: since the paint is red, it isn't yellow, it's either red or green, and at least something is red. But the paint's being red isn't equivalent to any of these other facts, because any or all of them might have obtained even had the paint not been red.

Evidently, non-atomic contents cannot in general amount to particular combinations of atomic contents: since different incompatible combinations can equally well satisfy a given representation, the structure represented must be conceived more abstractly. One attractive suggestion is that the represented content is a structure or feature of the whole world that it shares with various alternative possible worlds—for instance, its belonging to a certain set of such worlds. For our purposes, however, it's irrelevant how exactly this might go; and the real lesson lies in why it's irrelevant. Insofar as *any* representation represents something about the world, and can misrepresent it, the content of that representation can be conceived in terms of a partition of possible worlds. Thus, the content of a certain map might be associated with the set of worlds of which that map is accurate. But this shows that whatever distinguishes some representations as logical gets lost if content is reduced simply to a partition of possible worlds. Or, to put it the other way around and more broadly, whatever in the content distinguishes the genera must lie in distinct general *sorts* of "possible" world. Sets of compossible atomic facts, for example, might plausibly be the worlds needed for logical content, whereas totalities of compossible shapes of terrain or compossible scenes might amount to possible "iconic worlds". But what underlies *these* differences?

The following, I believe, is the heart of the matter: the primitive elements of logical contents—whether objects and properties, components and complexes, conditions and actions, or what have you—are always identifiable separately and individually. That is, they can enter into atomic contents one by one, without depending on their concrete relations to one another, if any. It can be a fact, for instance, that the glass is broken, quite apart from whether it is larger than the cockroach, contains traces of whisky, or used to belong to the deceased. The point is not that all combinations are possible—obviously the glass can't be both broken and intact—but rather that it is not a precondition on its being broken (or representable as such) that it or its brokenness occupy any determinate position in any structured ensemble of other objects or properties; they stand in their togetherness on their own. Seen from the side of language, this self-standingness of fact elements shows up in the mutual independence of proper names and the (in this regard) namelikeness of primitive predicates. For reasons that will emerge shortly, I call these self-standing logical content elements (such as individual objects and properties) *absolute* elements.[12]

The contents of iconic representations are in general quite different. Taking a cue from graphs, and remembering the quip about photographs, iconic contents might be conceived as variations of values along certain dimensions with respect to locations in certain other dimensions. Thus, variations of temperature with respect to time, or altitude with respect to latitude and longitude, would be the contents of familiar icons. The former dimensions are called dependent, the latter independent, because, for each point in the relevant independent region, a point is determined in the dependent space, but not necessarily vice versa. The (skeletal) contents of a color photograph have two independent and three dependent dimensions; monochrome photos reduce the latter to a single dimension, and line drawings and silhouettes further restrict that dimension to two (or a few) points. Representational sculptures can be thought of as silhouettes with three independent dimensions; and scale models are much the same, but with more internal detail. Analog computers typically represent the variations of quite a few dynamic variables, all with respect to time, or time and space.

This account, however, is insufficiently specific, until the notion of dimension is further delimited. Consider, for a moment, some unstructured sequence of distinct objects as a "dimension"; and imagine a space with *n* such independent dimensions, and one two-valued dependent dimension. Then, for any *n*-adic property, a "graph" in that "space" could show exactly which *n*-tuples of those objects have that property. Pressing the idea to its limit: if all the primitive objects in some universe of discourse were in such a sequence, and there were as many independent dimensions as the highest "adicity" of any primitive property, plus a separate dependent dimension for each primitive property, then the entire object/property state of that universe could be represented in a single multi-dimensional silhouette. What's wrong with this "picture"? The problem lies in regarding an arbitrary sequence as a dimension.

On a genuine *dimension* (in the definition of iconic representation) the relative positions of the content elements are not at all arbitrary. Quite the contrary, those elements are always organized relative to one another in some regular structure that every representational token of the pertinent scheme essentially presupposes (and which, therefore, cannot itself be represented by any such token). For instance, both dimensions corresponding to a graph of temperature versus time are uniform measurable magnitudes that admit of simple metrics; and the very possibility of the graphing scheme depends on this. Why? Because what is graphed is the *shape* of the variation; hence, no particular temperature or instant is relevant except as placed, along their respective dimensions, relative to all the others. The point is even more conspicuous for maps and pictures: the shapes of the variations are the represented (skeletal) contents—and "shape" makes no sense except in terms of relative locations in structured spaces. With obvious and I hope illuminating contrast to the above use of 'absolute', I call the dimensionally placed elements of iconic contents *relative* elements.

The point is not that the dimensions are continuous. Thus, for a biological species that reproduces synchronously, it makes perfect sense to graph population against generation, even though both variables are discrete. The graph makes sense because generations and population sizes are themselves both intrinsically ordered. This does not mean merely that they can be assigned numbers; rather, what matters is the order in the dimensions themselves—an order that can be captured in an assignment of numbers, if desired. By contrast, a "graph" of telephone numbers versus social security numbers would be nonsense, precisely because those assignments do not reflect any underlying order. (They are, in effect, just absolute names, formed from digits instead of letters.[13]) It is also not necessary that the dimensions (or spaces), either dependent or independent, be Euclidean; they might be, for instance, curved or ramified, so long as their intrinsic structures enable relative locations and hence representable shapes.

The contents of logical representations are not, in any analogous sense, "shapes" in "conceptual space"; and the reasons show up the limitations of the spatial metaphor for factual possibilities. Only spaces in which locations are identifiable with respect to intrinsically ordered dimensions—such that relative position makes non-arbitrary sense—support a notion of *shape* at all. Objects and properties (or kinds) do not as a rule constitute such dimensions: one cannot place Mugsy "between" Toto and Lassie on the dog scale, or spaniel "above and to the left" of collie in breed space. Therefore, even if for every dog there were a breed, it would make no more sense to graph breed against dog than phone number against social security number: there's no space for such a graph to take shape in. Objects have their properties one by one, absolutely, and not as part of some shape, relatively.

Insofar, however, as contents are worldly, it may seem that logical and iconic contents can overlap or coincide. We can, for instance, say—it is a fact—that the Earth is round. Is this not the very same "structure or feature of the world" that would be represented also by a silhouette of the Earth against a bright background? I don't think so. In the first place, the sentence identifies (the fact comprises) a particular object and a specific property of that object, whereas the silhouette identifies no object or property: its skeletal content is just the overall pattern of light and dark from some perspective. But further, though the sentence is entirely compatible with the Earth being transparent or just as bright as the background, the silhouette is not; and so on.

The thesis that representational genera are distinguished according to the structure of their contents yields an unexpected dividend: it explains and therefore supports the observation made earlier that "translating" from one genus to another requires wits. If the skeletal contents of two generically different representations—say, a picture and a description of the scene of the crime—differ qualitatively, even in their basic structure, then, in particular, neither includes the other. That is, much (or all) of what the one representation represents is simply not represented at all by the other. Thus, a description of a situation does not "say" how the light values vary with angle of view, any more than a photo of those values "graphs" what objects are present with what properties. Hence, a witless conversion is not possible, simply because the content is not there to convert. On the other hand, a system with wits of its own—background familiarity with the world and the circumstances—might be able to "tell" or to "see" what that missing content would have to be, and fill it in. Often, for instance, a person, relying on knowledge and experience, could tell that a certain light pattern would normally issue only from an object with a certain property, and could thereby supply the logical content needed to "translate" an image into words.

8 Connectionist networks

Is it possible to give a comparable generic account of the contents of distributed representations? In contrast to the familiar logical and iconic genera, distributed representation is a relatively new idea, and remains largely terra incognita outside the circle of scientific specialists. So far, acknowledged examples tend to be arcane, and disputes continue as to which of their features are most significant. What's more, the distinctions between recording and representing and between skeletal and fleshed-out content, no matter how clarifying in the long run, can only add perplexity at the outset. Accordingly, it will not be possible in this case, as before, to draw out a broad characterization from a few intuitively paradigmatic examples, relying on the reader to appreciate how the generalizations work. Instead, therefore, I will attempt a concise overview of a central class of cases, and then propose a generic account that seems both parallel to and interestingly different from the logical and iconic genera considered above.

The most widely investigated examples, and those with regard to which the term *distributed representation* was first introduced, are connectionist networks. Though there are many species, the following characterization applies to a representative variety, and gives some indication of how they can differ.

A network consists of a large number of independent elementary units, each of which has a variable state, called its *activation* level.[14] Each unit also has a number of inputs and/or outputs (via connections—see below), and corresponding *transfer functions* relating these to its activation level. Often, the activation is determined as a function (possibly stochastic) of its previous value and the sum of all the inputs; and the output is a function of the current activation. Both functions tend to be simple, though they are seldom both linear. Usually, the transfer functions of all units are the same (or of just a few distinct types). More complicated arrangements, functions, and states are, of course, possible; but they have the effect of locating more of the system's "power" in the individual units, and less in the network—contrary to the spirit of connectionism.

The units of a network have a great many *connections* among them, each of which has a determinate strength, called its *weight*. As a rule, each connection connects one "source" to one "target" unit, with no two having the same source and target. In the most general such case, each pair of units is connected in each direction, and each unit is connected to itself (so the number of connections is the square of the number of units). Many networks, however, implement only a subset of these connections; and they can be classified according to which connections are included. The most significant distinction is between *cyclic* and *acyclic* nets: in the former, but not the latter, sets of connections can form closed paths (feedback loops). The general case is cyclic. By contrast, *feed-forward* networks, in which the units are effectively subdivided into two or more "layers" by implementing only connections from one layer to the next, are acyclic. (Perceptrons are two-layer acyclic nets.)

The input to a network as a whole consists of a pattern of initial activations assigned to some (perhaps all) of its units. The output consists of the final activation levels of some (perhaps the same, perhaps different) units. These final values are reached during a "run" in which the units individually adjust their activations in response to incoming values, and then broadcast outgoing values, according to the respective transfer functions. The incoming value from a connection to its target generally equals the product of that connection's weight and the value that was broadcast by its source. Thus, the output of a network is a joint function of its input and the weights on all its connections (plus, perhaps, some random factors). In the case of a feed-forward network, the input is entered on the first layer, and then each successive layer is activated, until the output appears on the last. Cyclic networks are more difficult because the feedback can generate instabilities (loops that feed on themselves chaotically); hence, they must be carefully constrained to ensure that they will settle down and yield an output.

Corresponding to this broad characterization of the networks themselves, we need a comparably broad characterization of their performance. In general, a network's inputs and outputs are both *patterns* of unit-activations: given a certain input pattern, the net will produce some associated output pattern. What makes this different in principle from a traditional look-up function (besides efficiency) is the network's fundamental tendency to group input and output patterns according to *similarities*: that is, similar inputs tend to yield similar outputs. One can get a rough grip on why this should be so from information theoretic considerations. Loosely speaking, the number of possible input and output patterns is exponential in the number of nodes, whereas the number of connections is only polynomial; so there aren't enough "bits" in the connection strengths to encode an arbitrary distinct output pattern for each possible input pattern.[15] Rather, arbitrary output patterns can be encoded for a number of relatively scattered input patterns, leaving the rest to fall where they may. Of course, the network will produce some output or other for any given input; but these associations must be more or less determined by those already encoded—specifically by "clustering" around them in some way.

Importantly, there can be multiple independent respects of similarity in play at once. Thus, a given input might be classified as belonging to (similar to) one group of possible inputs in one respect, and, at the same time, as belonging to an entirely different group in some other respect. Operationally, of course, the groupings themselves define the "respects": for two patterns to be "similar" is nothing other than for them to belong to the same cluster. The reason that the groups can be thought of as kinds (similarity groups) is that a network can "learn" them from examples, and then generalize to new cases. Depending on the training set and other details, the generalizations can be both surprising and strikingly apt. (The teachability of networks has also been important to connectionist research for a more workaday reason: trivial cases aside, there is no known practical way to determine suitable connection weights for given problem domains other than via automated training regimes.)

Different networks induce different groupings, hence different respects of similarity. How does a network induce a similarity grouping? For the simplest classifier, a domain is partitioned as follows: two inputs belong to the same group just in case they produce (exactly) the same output. But this is doubly restrictive: first, it does not allow for multiple respects; and second, it does not take account of outputs being "almost" the same—that is, similar. Both restrictions are lifted in principle by letting the outputs of one network be the inputs of one or more others. If the latter are simple classifiers, each can define an independent respect of similarity in the outputs of the first network, which can in turn induce subtler simultaneous respects of similarity in its inputs. Indeed, still another network, connected to the outputs of all the networks connected to the first, can pick up patterns in the co-occurrence of simultaneous groupings, thereby inducing still more sophisticated ("meta") groupings in the original input patterns. But all of this is just to indicate indirectly the potential power of multi-layer (or cyclic) networks.

Yet even this fails to capture the possible richness in the input/output behavior of a sophisticated network, for it ignores the potential for relevant similarity relations among the output patterns. The point is not simply that outputs can significantly cluster—though that's important—but that relative positions of output patterns within their similarity groups can systematically reflect the positions of the input patterns within their similarity groups. Thus, to take a contrived but suggestive example, imagine the input and output patterns respectively coming from and going to a ping-pong player's eyes and muscles. Major similarity clusters in the input will effectively "recognize" slams, slices, and lobs in the opponent's stroke, as well as (simultaneously) topspin, bottomspin, and lateral English, not to mention speed, direction, and the like, in the oncoming ball. Output patterns must effect major partitions among forehand, backhand, and scoop shots, and upward versus downward wrist motion, plus, of course, locating and aiming the racket. Not only must the output "choices" be coordinated in subtle ways with the input recognitions, but also—and this is the point—everything depends on finely tuned gradations within those major similarity groupings: variations and nuances in the arriving shot call for *corresponding* variations and adjustments in the return.

9 Are there "distributed" representations?

As examples like these illustrate, connectionist networks contain two quite different candidates for status as representations. On the one hand, there are *patterns of activation* across various sets of network units—the input and output sets, obviously, but also perhaps others, such as intervening layers. So, in the ping-pong example, patterns might represent any number of concrete and abstract aspects of ball and racket motions, gathered and/or projected. On the other hand, there are *patterns of connection weight* among the units of the net. It is usually much more difficult to say what these might represent; but it is somehow they that encode the distinctive capabilities of each individual network. Thus, continuing the example, it would be the connection strengths (collectively) that "contain" the system's know-how or ability to play ping-pong. In contrast to patterns of activation, which are ephemeral (changing with the play in real time, say), connection weights tend to remain constant, or to change only slowly (with experience and practice, say).

Both sorts of pattern are commonly cited as instances of distributed representation. But in most actual research systems, input and output patterns tend to be, at best, distributed *recordings* of various determinate domain features, structures, or categories—in effect, lists or descriptions. Thus, either individual activations indicate particular features separately (not distributed at all, but "localist"), or else the representations of a number of features are spread out and superimposed on a multi-unit pattern of activation—distributed in the recording, but still, so to speak, localist in content. There is a perfectly sensible reason for this: investigators must be able to interpret inputs and outputs if they are to argue that their systems are representational at all, and to evaluate their performance. Inasmuch as descriptive (logical) representations are the most familiar, it is natural to want to interpret others as equivalent or comparable to these—especially when it is taken for granted that the generic differences lie not in the represented contents but in the representing relation. Activation patterns of hidden (non-input/output) units are often less fathomable; but even here the temptation is to look for "micro-features" or "coarse-coded" partitions.

Patterns of connection weight, on the other hand, are almost invariably quite undecipherable. These constitute not occurrent or episodic states that arise and pass away in the course of the net's operations ("thoughts"), but rather its long-term functionality or competence ("know-how")—which it is essentially impossible to express in words, except as generalities. But the fact that the best verbal characterization of a skill is an imprecise hand-wave does not mean that the skill itself lacks delicacy and detail; on the contrary, it has a "complexity" and "precision" of its own. To the extent, therefore, that connection patterns are regarded as representational at all, they are the most compelling candidates in contemporary networks for illustrating a *distinctive* representational genus—to wit, *distributed representations* (as opposed to mere distributed recordings).

Does it even make sense to regard the embodiments of a system's abilities or know-how as *representations*? Why not take them rather as just complex dispositional properties—acquired and subtle, perhaps—but, for all that, no more representational than a reflex or an allergy? Recall the characterization of representations as such adumbrated at the outset: they serve as a kind of "stand-in" for specific features or aspects of the environment to which the system must adjust its behavior, even on occasions when those features or aspects are not currently present or detectable. And consider, in those terms, a very special but important case of ability or know-how: the ability reliably to recognize the individual faces (or voices, or smells, ...) of one's regular companions. It seems to me to be unquestionably some sort of structure or feature *of the environment* that this face can look all these different ways (different angles, expressions, surroundings, and the like), another face can look all those other ways, and so on, with hardly ever any overlap. Clearly, such a "feature" could never be detected on a given occasion; yet "adjusting" to it *in absentia*, as a means to correct reidentifications, would be of great value. Accordingly, whatever incorporated the ability to recognize those faces could, by our account, be deemed a *representation* of that feature. To be sure, that initial account was formulated only casually and without argument, so nothing is proved by it; still, the consilience is not unsuggestive.

In particular, it suggests a natural generalization to other abilities and know-how. Thus, it is likewise a structure or feature *of the world*—quite determinate and intricate, albeit in some sense also highly "abstract"—that just these patterns of movement will suffice to get those ping-pong balls back over the net, a car through this traffic, or, indeed, that fresh antelope home to the cubs. Again, such structures are never presently detectable, yet adjusting to (taking account of) them in real time would be just as valuable. Surely, in fact, this separation of recognitive and performative abilities is altogether artificial: special cases aside, skillful performance is essentially adaptive to current circumstances, which is to say responsive to recognized relevant developments; and recognition in general can be nothing other than production of responses relevant to what is recognized.

Representations as such, however, must be party to a general scheme. How do patterns of connection weights, understood as representations of the subtle worldly structures accommodated by abilities and know-how, belong to a scheme? There is certainly little, if any, prospect of finding explicit interpretation or projection rules that would specify, for each "well-formed" weight pattern in some net, what all it would represent. But this is at most to acknowledge that schemes of distributed representation may not be articulately definable. It remains the case that a network, incorporated in a real-world system (an organism, for instance), and typically encoding a considerable variety of complicated responsive dispositions, *could have* encoded any of an enormous range of others, if only its connection weights had been different. Moreover, for that system, the actual weights consistently *determine* (fix) which abilities are actually encoded; whereas for a (qualitatively) different net or system, those weights might determine quite different dispositions, or none at all. Finally, there are clear possibilities of malfunction or malperformance in the reliance upon and/or management of actual weight patterns: for any of many reasons, a system can misrecognize or misbehave, even within the range of its normal abilities; and weight modifications in the light of experience (learning) can be carried out improperly, or result in degraded performance. Thus, whether an explicit "semantics" is possible or not, it does seem that weight patterns can be regarded as belonging to representational schemes.

The "elements" of the contents of distributed representations, at least as we have considered them so far, would be the various aspects or looks of recognizable individuals or kinds and the various behaviors or movements of skillful responses—that is, the elements of competent situated action. What distinguishes these essentially from the absolute and relative elements of logical and iconic contents? The answer must lie in how they belong together, in two interdependent ways. First, such elements can belong together in that they fall in the same similarity cluster—for instance, by being aspects of the same individual or executions of the same move. Second, they can belong together in that one calls for the other, or determines it as relevant or apt—for instance, when a recognized eventuality calls for a certain response. These are interdependent because: (i) it is by virtue of calling for similar responses that elements belong to the same recognition clusters, and vice versa; and (ii) it is by virtue of belonging to a certain recognition cluster that an element can call for a certain response, and vice versa. With due, and I trust understandable, trepidation about the history of the term, I call elements that belong together in this twofold interdependent way *associative* elements.

It is perhaps worth mentioning explicitly three respects in which this account of distributed representation differs essentially from classical associationism, such as Hume's. First, and most conspicuously, what are here deemed associative are elements not of representations (ideas, for instance) but of represented *contents* (structures or features of the represented world). Second, by traditional lights, the mind associates ideas because they are antecedently ("objectively") similar or otherwise related; whereas here the similarity of elements is consequent on ("induced by") a system's know-how or ability effectively casting them into the same clusters. (Note carefully: this is not incompatible with it being a structure or feature *of the environment* that such and such intricate responsive dispositions constitute competence in it, relative to given ends.) Third, and hand-in-glove with the preceding, traditional theories did not understand the different associative relations, especially resemblance and constant conjunction, as interdependent; but we have seen that clustering together and calling for a certain response make sense only in terms of one another.

10 Conclusions and morals

The upshot of these considerations—tentative and exploratory, I reiterate—is, first, that representational genera are distinguished essentially by the characteristic structures of their represented contents; and, second, that the represented contents of logical, iconic, and distributed representations are structurally characterized respectively by absolute, relative and associative elements.

In conclusion it may be observed that such an account of the genera, to the extent that it is tenable, yields another unanticipated dividend: by, in effect, dividing the spoils, it may mitigate the PDP/GOFAI standoff in contemporary cognitive science. Thus, consider the following putative dilemma: either PDP networks "implement" familiar symbolic AI architectures (in which case they are *irrelevant* to cognitive theory), or else they are incapable of representing contents which we know to be cognitively representable (in which case they are *inadequate* to cognitive theory). But if it is acknowledged at the outset that logical and distributed contents differ fundamentally in kind, then proponents of network models can cheerfully embrace both horns of the alleged dilemma, without conceding any damaging consequences. For, on the one hand, people obviously speak, and can think what they can say. (And no doubt that facility is implemented in some sort of network: what else, if the brain is a "neural net"?) But that doesn't show the cognitive irrelevance of network models, unless it is further assumed that logical contents are the *only* contents important to cognition; and there's no reason to assume that. On the other hand, acknowledging the qualitative difference in kinds of content is already to acknowledge that not all contents can be contents of distributed representations. But this implies no inadequacy of network models unless it is further assumed that distributed contents are the *only* contents they can allow as important to cognition; and there is no reason to assume that. Therefore, the camps are not really at odds, unless one (or each) purports to account for everything cognitive—but there's no reason to put up with that.

Indeed, the relationship may be considerably more intimate. Logical contents require absolute elements, such as determinate, reidentifiable particulars and properties. But among the most frustrating unresolved problems of "good old fashioned" AI and philosophy of mind is how these contents are recognized in realistic situations. (Constructing explicit tests in terms of neutral sense data, for instance, has proved quite unrewarding.) *Recognition*, however, is a specialty of systems employing distributed representations—suggesting the possibility of a deep cooperation, based on an *ability* to apply names. But any partnership would be unlikely to stop there, especially insofar as language is an evolutionarily late capacity, implemented in organs originally adapted for another sort of representation. Thus, everyday speech is extraordinarily sensitive to relevance and topical surprise, yet remarkably unfazed by ambiguity, ungrammaticality, and catachresis; it is thoroughly suffused with allusion, trope, posture, and drama; it is biassed, fanciful, opportunistic, emotion charged, and generally fast and loose; and much the same could be said of thought. Yet these qualities are notoriously resistant to capture in explicit symbolic models. Perhaps, that is because they involve adjustment to contents that cannot be represented logically: sophisticated and intricate abilities that, in symbol processing terms, could only be classified as transductions or basic operations—but are hardly peripheral or primitive. Perhaps natural language is possible only as symbiotic with the distributed representations of the system it is implemented in.

Notes:

[1] This use of the term 'content' is not altogether standard. Most contemporary authors (and I, in the other essays in this volume) mean by the "content" of a representation something distinct from the object it represents, and which determines that object (as sense determines referent, for instance). Here, however, I mean by 'content' that which the representation represents—the "object" itself—but *as it is represented to be* (whether it is that way or not). Thus, it is a possible object—which may in fact be actual, or similar to something actual, or neither. [Note added 1997.]

[2] For instance, if (or to the extent that) particular representations are tokens of well-defined types, the scheme will determine the content of any given token as a function of its type—or, at least, these will determine how that content is determined. Thus, if any extra-schematic factors (such as situation or context) co-determine contents, then which factors these are and how they work are themselves determined by the scheme and type.

3 "Without loss" here means without loss of representational content; it goes without saying that there could be other sorts of loss, such as market value (think of a digitized Rembrandt) or convenience (a digital recording often cannot be read, heard, or viewed without special equipment).

4 The above definition is essentially that offered by Hinton, McClelland, and Rumelhart 1986, 77; and by McClelland and Rumelhart 1986, 176, but in a less binding terminology. The two-volume set—the "PDP volumes"—in which both of these essays appear has become something of a *locus classicus* for research on parallel distributed processing. The most extended and penetrating treatment I have seen of distributed representation as such is by van Gelder (1989), who argues persuasively that superposition is more fundamental than spread-out-ness. (I have retained the more usual account merely for expository convenience.) Philosophical implications of connectionism are also discussed in Horgan and Tienson 1987; Pinker and Mehler 1988; and Churchland 1989. Note, incidentally, that the notion of separate tokens for distinct contents is undermined by distributed representations, to the extent that there is just one "big" token representing many different contents at once.

5 This procedure would be anything but ridiculous if the inscription happened to be in an unknown script, such as on the wall of an ancient tomb. Because the photograph will preserve all visible features of the inscription, it will *a fortiori* preserve all representationally relevant visible features, even if no one yet knows what they are.

6 Such graphical displays of points in weight-space (sometimes called "Hinton diagrams") are common in the connectionist literature; see, for example, Hinton, McClelland, and Rumelhart 1986, 103.

7 Fourier transforms of spatial intensity functions, expressed algebraically as sufficiently long sums of terms, would be comparable logical recordings of distributed representations.

8 A general procedure here means an effective procedure that does not depend on there being an upper bound on the number of distinct representational types in either scheme. If there are such

witless transforms in both directions, the relation between the schemes may have to be even closer—perhaps as close as being "notational variants" of one another.

9 Note that witlessness is not a function of level of description. Thus, it is sometimes said that whether certain processes in the brain (or a computer) count as intelligent or sensible depends on the level at which they are described; the very same processes can be intelligent at the cognitive level, but simultaneously unintelligent and mechanical at a lower formal or physical level. Be that as it may, witlessness here means *thoroughly* witless—there is *no* level at which the process can be redescribed as other than oblivious to content and ignorant of the world

10 "Internal" here (and hereafter) means *functionally* internal, not "immanent to the sphere of consciousness", or any such epistemological/metaphysical status.

11 In general only *possible* because the sentences might be false; "actual" facts would be the contents of true sentences.

12 Two qualifications. First, to call objects and properties (or whatever) "absolute" is not to imply that they are what they are independent of the scheme in terms of which they are represented; that is, the account of logical content in not meant to bear on questions of realism. Second, the self-standingness at stake is what might be called "occasion", as opposed to "constitutive" self-standingness. Thus, it may be constitutive of the kind 'spaniel' (as identifiable in terms of some scheme) that it be a breed of 'dog' distinct from 'collie' and subsuming 'cocker'. But, on any particular occasion, whether some object is of some kind is a function just of that object and kind, and not of any determinate relationships in which either may stand (except, of course, to the extent that the kind is itself relational).

13 The point is not simply that there is no interesting correlation between phone number and social security number, as there might be no interesting correlation between blood pressure and rate of hair growth, but rather that there is nothing there to correlate. Of course, the numbers do identify particular telephones and people, respectively; but they do not locate them

along any intelligible dimensions in ways that might or might not be correlated.

14 Activations can be discrete, real, or (presumably) even complex valued, bounded or unbounded. Units can also have additional state variables, such as latency or fatigue conditions, or a kind of momentum in the activation itself.

15 The argument is "hand-wavy" in various respects, most notably in effectively assuming that the activation values and connection strengths are both only finitely precise (and to approximately the same degree).