Varieties of Two-Dimensionalism

Jim Pryor
Princeton University
2/10/03

There are different *kinds* of two-dimensional matrix one can work with, representing different properties of an expression. One has to understand the rows and columns differently for the different matrices; but there are some formal characteristics all the matrices have in common.

The Core of the Formalism

Suppose we're assigning a two-dimensional matrix to an expression E. The cells are always filled with an entity of a sort that's suitable to be an extension for E. So if E is a sentence, the cells are filled with truth-values. If E is a singular term, the cells are filled with objects. And so on.

Along the top of the matrix, we label the columns with what I'll call "Points of Evaluation." For most kinds of matrix (though not all), these will be metaphysically possible worlds.

Down the side of the matrix, we label the rows with what I'll call "settings." Each setting "generates" a partial mapping from Points of Evaluation to extensions; that will be the relevant row in the matrix. For different kinds of matrix, we'll have different stories about what the settings are, and how they generate these rows.

Each setting also counts as "belonging to" exactly one of the Points of Evaluation. The function from settings to the Points they belong to may be many-one, and need not be onto the set of Points.

To illustrate, here is a matrix for a sentence S. On each row, I've shaded in the cell corresponding to the Point of Evaluation that that row's setting "belongs to."

	Point of Evaluation 1	Point of Evaluation 2	Point of Evaluation 3	…
Setting 1	Each row is the mapping (from points of evaluation to extensions) "generated by" that row's setting.			
Setting 2	Since this is a matrix for a sentence, cells are filled with truth-values.	*true*	*undefined* (the mapping from points to extensions may be a partial function)	*false*
Setting 3		Setting 3 "belongs to"		

		Point of Evaluation 2		
Setting 4		Setting 4 also "belongs to" Point of Evaluation 2		
Setting 5 …			Setting 5 "belongs to" Point of Evaluation 3	

Consider the function from settings to the values in the shaded-in cells. (So, this function will map setting 2 to false, will map setting 3 to whatever truth-value is in the shaded cell <setting 3, Point of Evaluation 2>, will map setting 4 to the truth-value in the shaded cell <setting 4, Point of Evaluation 2>, and so on.) This function will have a special role in the philosophical use of these matrices. It goes by different names. Sometimes it's called "the primary intension" of the matrix. Sometimes it's called "the diagonal intension." Jackson calls it "the A-intension." It's not important for present purposes why these names were chosen. I will just call it "the matrix's diagonal." (This is an awkward name, since the shaded-in cells need not really form a diagonal. In general it won't be true that setting *n* belongs to Point of Evaluation *n*. But we'll go with this name because it has gained some currency in discussions of two-dimensionalism.)

The rows in the matrix also go by a variety of names. They've been called: "the horizontal intension," "the secondary intension," and "the C-intension." I will just call them "rows," or more fully, "the horizontal row for setting *n*."

Here's a toy example. Understand the points of evaluations to be possible worlds. Understand the settings to be world-bound philosophy conferences. That is, each philosophy conference takes place in exactly one possible world. The setting "belongs to" the world in which it takes place.

For a given sentence S, here's how we fill in the rows. We go to the relevant philosophy conference for that row. We look for the first utterance of S at that conference. If there is such an utterance, we let the row be the intension expressed by that utterance of S, in the language it was uttered. (At different conferences in different possible worlds, S will be uttered with different meanings.) If there was no utterance of S at the conference, then we instead fill in all the cells with "undefined."

The diagonal for this matrix will then be a function from world-bound philosophy conferences to the truth-value, in the world of the conference, of the first utterance of S at that conference, in the language it was uttered.

Of course, *that* diagonal function will not be useful for very many philosophical purposes. I've chosen this example because it illustrates how much flexibility we have in defining what the settings are, how the rows are to be generated, and so on. Depending on how we settle those questions, we'll get a variety of different matrices. And these will not all be useful for the same philosophical purposes.

Different Matrices and Their Philosophical Interpretation

The most familiar use of this kind of formalism is in Kaplan's semantics for indexicals.

In a **Kaplanian Matrix**, the points of evaluation are possible worlds, and the settings are "contexts of utterance" relative to which we can interpret a sentence. (Some theorists require these contexts to be situations in which a speaker genuinely is uttering the sentence. Other theorists, like Kaplan, do not require this.) These contexts of utterance are understood to be located in exactly one world. Kaplan says that every sentence has associated with it a *standing meaning* or *character*, that, relative to a given context of utterance, yields a proposition. If the sentence contains indexicals, then its standing meaning will give us different propositions for different contexts. For instance, the standing meaning of "I can juggle" will give us the proposition *Jim Pryor can juggle* relative to contexts in which I am the agent, the proposition *Clio can juggle* relative to contexts in which Clio is the agent, and so on. Each row in the matrix corresponds to the intension of the relevant proposition.

Now some indexical sentences have meanings with the following interesting property: relative to any context of utterance, that meaning yields a proposition which is true in the world of that context. For instance, "I exist" and "I am here now" each have that property. Those sentences could of course be given different meanings in other possible worlds. But their *actual* meaning is such that, relative to every context c, it yields a proposition which is true in the world of c. (The agent of c will always exist in the world of c; and in that world, will always be present at the time and place of c.) Kaplan calls this property *indexical validity*.

For theorists who, unlike Kaplan, define contexts of utterance to be situations in which the agent genuinely is uttering sentence S, then "I am speaking" and "I am referring to myself" will also be indexically valid.

The diagonal of a Kaplanian matrix will serve as a yardstick of whether a sentence is indexically valid. The sentence will be indexically valid iff its diagonal maps every context to true.

Kaplan believed that whenever S is indexically valid, "I can know *a priori* that S" will be true. This claim is controversial; and my own view is that it's false. However, it is an early example of a claim that there is some link between *a prioricity* and having a diagonal that maps every context to true. That is the kind of thing that two-dimensionalists are aiming for. They want to employ some kind of matrix for "water" that looks like this:

	World #1	World #2	World #3
Setting #1, in an environment in world #1 where H_2O is the local clear drinkable liquid	H_2O	H_2O	H_2O
Setting #2, in an	XYZ	XYZ	XYZ

environment in world #1 where XYZ is the local clear drinkable liquid			
Setting #3, in an environment in world #2 where XYZ is the local clear drinkable liquid	XYZ	XYZ	XYZ

They want the matrix for a sentence like "Water is a liquid" to look like this:

	World #1 (H_2O and XYZ are both liquids in their respective environments)	World #2 (H_2O is always in solid form, XYZ is a liquid)	World #3 (H_2O and XYZ are both always in solid form)
Setting #1 (H_2O)	true	false	false
Setting #2 (XYZ)	true	true	false
Setting #3 (XYZ)	true	true	false

The diagonal of that matrix maps every setting to true; this corresponds to the fact that it's knowable *a priori* that water is a liquid. In every setting in which you use the expression "water"—be it an Earthian setting or a Twin-Earthian one—you end up referring to *something* that's a liquid, in the world of that setting. Sometimes it's H_2O, other times it's XYZ, but it's always a liquid.

Matrices for sentences S where "It's knowable *a priori* that S" is false, on the other hand, should have diagonals that map some settings to false. For example, the matrix for "Water contains hydrogen" should look like this:

	World #1 (in every world, H_2O contains hydrogen but XYZ does not)	World #2	World #3
Setting #1 (H_2O)	true	true	true
Setting #2 (XYZ)	false	false	false
Setting #3 (XYZ)	false	false	false

This diagonal maps some settings to true and others to false; that corresponds to the fact that it's knowable only *a posteriori* that water contains hydrogen. In some settings in which you use "water," you refer to a substance that does contain hydrogen (H_2O), in others you do not. And you can't know *a priori* which of those settings you're in.

That's the kind of thing that the two-dimensionalist wants. It's not clear whether he can get it. We need to look more closely at how these matrices are defined, exactly.

Take some singular term T which refers in your language to an object o. We can assume that there will be some facts *in virtue of which* T refers to o. We can call these **the reference-making facts** for T. These facts will involve some complicated relationships between T, o, your linguistic intentions, the linguistic intentions of other members of your community, and other facts about your environment. Write those facts out like this: F[…T…o…you…your environment…]. Now take the relation of being a *t* and an *x* such that F[…*t*…*x*…you…your environment…]. It's because T and o stand in that relation that T refers in your language to o.

Now let's consider all the subjects in other situations, perhaps in other possible worlds, who understand or use expressions T* such that, in their world, there is an object that stands in this same F-relation to T*, and them and their environment, etc. In other words, $\exists x$ F[…T*…*x*…the subject of that situation…that subject's environment…]. Call any term T* which has these properties **the semantic counterpart**, in the relevant situation, of your word T.

Let's construct a matrix for T where the points of evaluation are possible worlds, and the settings are situations where a subject understands or is using some semantic counterpart of your word T. Each of these situations is understood to be located within a single world. We set the row to be the intension that the counterpart word has in the mouth of the setting's subject.

Call this **a Semantic Counterpart Matrix.**

The diagonal of such a matrix will be a function from settings to the referent of your word's counterpart in that setting. For example, if I use "this apple" to demonstrate an apple a1, and the subject in setting #8 uses "dwee arthxc" with the same standing meaning as my words "this apple," and he's using his words to demonstrate apple a2, then the diagonal for my expression "this apple" will map setting #8 to apple a2.

Now, one way in which a term can get to have a reference is if a speaker *stipulates* that it is to rigidly refer to whatever uniquely exemplifies some properties, or more generally, to whatever uniquely bears some relation R to his setting. (Note that we're taking R to be a relation, not the words that the speaker uses to specify that relation.) Suppose T is a term that gets its reference in this way. If we generate a Semantic Counterpart Matrix for T, then in every setting of that matrix, the subject of that setting will have *some* term T* that he stipulates is to rigidly refer to whatever uniquely bears R to his setting. Suppose o* is the object that does (in the world of the setting) uniquely bear that relation. Then (because T and T* are stipulated to be rigid), we fill in every cell on that row with o*. (Optionally, for worlds where o* does not exist, we can give the cell a special value of "*no reference.*")

For terms that get to have their reference **by this kind of reference-fixing stipulation**, we can also define a bigger matrix, that includes all the rows of the Semantic Counterpart Matrix, and some additional rows besides. The additional rows are ones where the subject *hasn't* made any counterpart stipulation. He needn't have any word that refers to the unique R in his setting. He need not even speak a language! Instead, we just

look directly to see what object o* *does* uniquely bear R to his setting; and we fill every cell in the row with that object. When there is no object that uniquely bears R to his setting, we put a "*no reference*" value in every cell.

We can call this **a Reference-Fixing Matrix**. The diagonal of this matrix will be a function from settings to whatever object uniquely bears R to that setting.

Note that we can only define a Reference-Fixing Matrix for expressions that *got* their reference by a reference-fixing stipulation. For expressions that got their reference in other ways, we haven't defined any way to generate the additional rows. For example, suppose Max sees a robot and forms the intention to call it "Otto." Suppose further that ostensive definitions of this sort *cannot* be reduced to any kind of tacit reference-fixing stipulation. Then we will be able to generate a Semantic Counterpart Matrix for Max's word "Otto"; but no Reference-Fixing Matrix. We've only defined what the rows look like for settings where the subject has an expression T* that he's introduced with the same kind of ostensive definition that Max used.

Now, two-dimensionalists like Jackson and Chalmers think that every case of ostensive definition—*and* linguistic deference, and every *other* way for a word to get a reference—can be reduced to a kind of tacit reference-fixing stipulation. So they think that *whenever* a word has a reference, we will be able to construct a Reference-Fixing Matrix. But this relies on their controversial views about ostension and deference.

So far, we've only defined Semantic Counterpart Matrices and Reference-Fixing Matrices for terms. The definition can extended in a natural way to sentences.

For the Semantic Counterpart Matrices, we require that every term and predicate in the original sentence have a semantic counterpart in the relevant setting. We set the row to be the intension had by the sentence made up of those counterparts, in the mouth of the relevant subject. For settings where the requisite semantic counterparts do not exist, the rows are undefined.

Reference-Fixing Matrices are an extension of this. For a Reference-Fixing Matrix, we require that each word in the original sentence *either*: (i) have a counterpart word in the relevant setting, *or* (ii) have gotten its reference by reference-fixing stipulation. We construct a proposition with the same structure as our original sentence, and which is populated by the objects and properties that are either the referents of the counterpart words, if there are such, or which uniquely exemplify the conditions that were used in the reference-fixing stipulations.

For example, suppose our sentence is "Julius loves Otto." "Julius" has gotten its reference by being stipulated to refer to whoever first invented the zipper (in this world). "loves Otto" has gotten its semantic content in some other way. Now let's go to setting #29, where the subject uses the words "luvkj ddwid" as semantic counterparts of your words "loves Otto." The subject in setting #29 may also—but need not—have stipulated that some word he uses refers to whoever first invented the zipper (in his world). Now our original sentence has the semantic structure <–object–, –property–>. We look for whoever *did* first invent the zipper in the world of setting #29. If there is no such person, every cell in row #29 gets the value "*contains a non-referring term*." But let's assume there is such a person, j*. He goes in the object place of our proposition; so we have <j*, –property–>. Next we take whatever property it is that the subject in setting #29

expresses with his words “luvkj ddwid.” Let that property be luv*. That goes in the property place. So now we have a complete proposition, <j*, luv*>. We set row #29 of our matrix to be the intension of that proposition.

Jackson and Chalmers want the 2D matrices *they* use to have an interesting epistemological property. They want them to be such that what you express by “I can know *a priori* that S” will be true iff the matrix for S has a diagonal that maps every setting to true.

Now, as I said, Jackson and Chalmers think that *all* our words get their reference by a kind of tacit-reference fixing. So they think we’ll be able to construct Reference-Fixing Matrices for any terms or sentences we understand. They *also* think that these Reference-Fixing Matrices will have the interesting epistemological property I just described. Both of these claims are quite controversial.

In some places, Chalmers defines a kind of matrix that has this epistemological property *by design*. Call this kind of matrix **an Epistemic Matrix**. I’ll explain how to understand these matrices. We’ll leave it an open question for now what correspondence, if any, there is between Reference-Fixing Matrices and these Epistemic Matrices.

For an Epistemic Matrix, the points of evaluation are no longer metaphysically possible worlds. Instead, we understand them to be **epistemic possibilities**. The notion of an epistemic possibility is an intuitive one, but it is hard to make rigorous. Still, many people want this notion to be clarified, and are optimistic about the prospects of doing so. So I think it’s fair for the two-dimensionalist to use it in his theorizing.

At a first pass, we can understand these epistemic possibilities to be sets of belief-types. Those are the beliefs that would be justified, if you learned that that epistemic possibility obtained. The epistemic possibilities should be internally consistent. They should also be consistent with some set of beliefs we assume you to have as background knowledge. For the two-dimensionalists’ purposes, this “assumed background knowledge” should consist of everything which you could come to know *a priori*, by reasoning in an ideal way. This set excludes many of your actual beliefs, even many beliefs you know to be true. It includes those of your beliefs you know *a priori*, those beliefs you now know *a posteriori* (or don’t know at all) but could come to know *a priori*, and a great deal of additional beliefs you do not now have. The epistemic possibilities are to be understood as supplementing this background knowledge. Hence, with the background knowledge I described, the epistemic possibilities will be ways that the world could epistemically turn out, consistent with what you can come to know *a priori*.

There are two further constraints the two-dimensionalist will want to put on these epistemic possibilities. First, they should be statable in some limited vocabulary: a vocabulary purged of any Twin-Earthable words like “water.” Second, they should be as maximally informative as they can, compatibly with the other constraints. So for every thesis that it’s possible to state in non-Twin-Earthable vocabulary, and that has a determinate truth-value, the epistemic possibilities should, together with the assumed background knowledge, *a priori* entail either that thesis or its negation.

OK, those are the epistemic possibilities that constitute our Points of Evaluation. The settings will be "centered" or *de se* versions of these epistemic possibilities. So if one Point of Evaluation describes the world as containing a human-shaped creature wearing a red hat, who is facing a dog-shaped creature wearing a blue sweater, and so on, then we will have the following *de se* epistemic possibilities:

- being a human-shaped creature wearing a red hat, facing a dog-shaped creature wearing a blue sweater
- being a dog-shaped creature wearing a blue sweater, facing a human-shaped creature wearing a red hat

and so on. These *de se* epistemic possibilities should also be statable in non-Twin-Earthable vocabulary, and otherwise be as maximally informative as they can. When combined with the assumed background knowledge, each *de se* epistemic possibility will be compatible with exactly one Point of Evaluation. That will be the Point of Evaluation that the *de se* possibility belongs to.

So now how do we generate the rows of an Epistemic Matrix? Let's begin with a rigid singular term like "Julius." Consider setting #15. Let's use the expression **the hypothetical knowledge associated with that setting** to include: what you know in virtue of understanding "Julius," together with your assumed background knowledge (i.e., all the things you could come to know *a priori*), and supplemented by the hypothesis that you're in the situation described by *de se* epistemic possibility #15. (Note that it needn't be *true in* epistemic setting #15 that you have any of this knowledge. Setting #15 might describe you as knowing very little.)

If this hypothetical knowledge is sufficient to identify one of the objects represented by it as Julius, then you should associate that object with every cell in row #15. That is the object you would conclude is the reference of "Julius," if you reasoned ideally and had the information that you're in the situation described by epistemic setting #15. (Modulo worries about setting #15 describing you as knowing very little.) If there is no such object, then every cell in row #15 is undefined.

(I said that when you've identified one of the objects represented by your hypothetical knowledge as Julius, you "associate" that object with the cells in row #15. In fact, there is a serious obscurity here. The objects that this hypothetical knowledge represents need not be real objects. They need not even be metaphysically possible objects. So one might fairly object to quantifying over them. Perhaps we should regard your hypothetical knowledge as a piece of fiction, and the object you "identify" as Julius as something like a fictional character, or an (incomplete) representation of an object. What we put in the cells of row #15 would then not be Julius himself, or any other object, but rather an (incomplete) representation of an object, one that according to the hypothetical knowledge associated with setting #15 exists and uniquely invented the zipper, and so on.)

Extending this formalism to non-rigid terms and to sentences is more difficult. Suppose that "the President of the U.S." is a non-rigid singular term (rather than a quantifier). If the hypothetical knowledge associated with setting #15 is sufficient to identify, not just:

- who is President in the Point of Evaluation that setting #15 belongs to

but also:

- who would be President in possible worlds that are as the other Points of Evaluation describe the actual world as being

then you set row #15 to be a function from the different Points of Evaluation to the people so identified. (The "serious obscurity" I noted above is present in this case, too.) When your knowledge is insufficient to identify anyone as meeting these criteria, then you leave the relevant cell(s) undefined.

Similarly for a sentence like "Julius is the President of the U.S." If the hypothetical knowledge associated with setting #15 is sufficient to ascertain, not just:

- whether Julius is the President in the Point of Evaluation that setting #15 belongs to

but also:

- what is the truth-value of the claim that Julius is the President with respect to possible worlds that are as the other Points of Evaluation describe the actual world as being

then you set row #15 to be a function from the different Points of Evaluation to the truth-values so ascertained. When your knowledge is insufficient to ascertain some of the truth-values, then you leave the relevant cell(s) undefined.

Suppose an Epistemic Matrix of this sort can be made sense of. Consider the diagonal of such a matrix, e.g., the diagonal for "Water is a clear liquid." Suppose for the sake of argument that if you understand this sentence, your hypothetical knowledge for every epistemic setting would enable you to know that it's true. That is, no matter which of the ways the world turns out, this sentence should, given your understanding of it, be true. Then for every setting, the sentence will get a value of *true* for the Point of Evaluation that setting belongs to. In other words, its diagonal will map each epistemic setting to *true*.

Next consider the diagonal for "Water contains oxygen." With respect to some epistemic settings, your hypothetical knowledge will enable you to ascertain that this sentence is true. Those will be settings where H_2O turns out to be the best referent for "water." With respect to other settings, e.g., ones where XYZ turns out to be the best referent for "water," your hypothetical knowledge will enable you to ascertain that "Water contains oxygen" is false. So the diagonal for this sentence will map some epistemic settings to *true*, and other ones to *false*.

Now "I can know *a priori* that water contains oxygen" is clearly false. And given the suppositions we were making, it looks like "I can know *a priori* that water is a clear liquid" should be true. After all, there's no epistemic setting you could be in, compatible with what you can know *a priori,* in which it's false. So we do seem to have a correspondence here of the sort the two-dimensionalist wants. "I can know *a priori* that S" comes out true iff the diagonal for S's Epistemic Matrix maps each epistemic setting to *true*.

Call any matrix that uses *metaphysically* possible worlds as Points of Evaluation, and centered metaphysically possible worlds as settings, a metaphysical matrix. All of the matrices we considered earlier were metaphysical matrices. These Epistemic Matrices are not. Now, many two-dimensionalists believe that they can construct metaphysical matrices that tightly correspond to, or that can do the same philosophical work as, an

Epistemic Matrix. Call this **the project of metaphysicalizing Epistemic Matrices**. I am very doubtful that this project can be carried out. I doubt that any metaphysical matrix can be defined whose diagonal will always have the same interesting epistemic properties as the diagonal of an Epistemic Matrix. However, I do not think that the project of metaphysicalizing Epistemic Matrices really needs to be a core commitment of the two-dimensionalist. Chalmers is attracted to the project, but he says at several places that he does not think the two-dimensionalist needs to commit himself to it.

What I Object To

There are some difficulties making sense of these Epistemic Matrices, which I have glossed over in my presentation.

In addition, as I said, I am doubtful about the project of metaphysicalizing Epistemic Matrices. But the two-dimensionalist can disavow that project.

My main objection to the two-dimensionalist project is this. Suppose that Epistemic Matrices can be made good sense of. That leaves it open how often we'll be able to construct Epistemic Matrices that have lots of defined cells. It could be that for *many* words and *many* epistemic settings, the hypothetical knowledge associated with that setting will be inadequate to identify an extension for the word. I believe that's how things actually stand.

(1) The two-dimensionalist believes that every term we understand gets its reference by a kind of tacit reference-fixing (or at least, that every term can be treated as if this were so). Hence, he thinks that for every term we understand there will be a well-defined Reference-Fixing Matrix. Most of its cells will have definite values.

(2) The two-dimensionalist also thinks there are interesting connections between Reference-Fixing Matrices and Epistemic Matrices. In brief, he thinks that if you fix the reference of "Julius" to be the inventor of the zipper, then in every epistemic possibility compatible with what you can know *a priori*, Julius invented the zipper (if he exists). The two-dimensionalist thinks this reference-fixing knowledge will enable you to identify the extension of your word, among the objects represented by most bodies of hypothetical knowledge.

That is why the two-dimensionalist believes that for every term we understand, he will be able to give definite values to most of the cells in its Epistemic Matrix.

I'll argue that (1) is false, and that (2) is also false. Hence, our reasons for expecting very fully-defined Epistemic Matrices collapse.