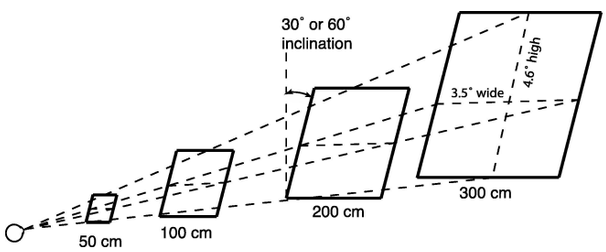
Vision: The Anti-Cognitive Revolution

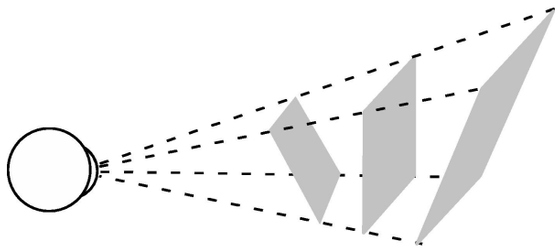
Paul Linton

The key question of perception for most theorists of perception is vision. And the key question of vision for most theorists of vision is 3D vision. So it's no surprise that 3D vision becomes the site of big battles over the nature of perception.

It's typical these days to describe vision as a "Cognitive" process. What this means is that the visual system is involved in "making a guess" about the outside world. For instance, when the outside world is projected on the retina, the resulting 2D retinal image is consistent with innumerable objects of different sizes and different shapes, and so vision has to "guess" what that true values of scale and shape are.

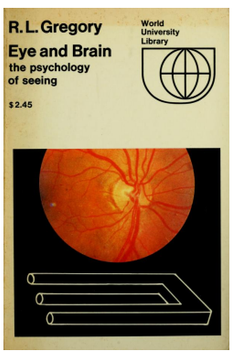


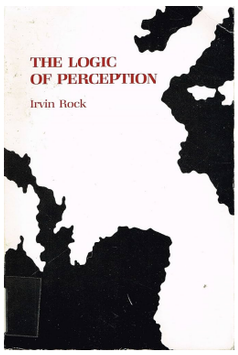
Ambiguity as to Scale (Size + Distance)

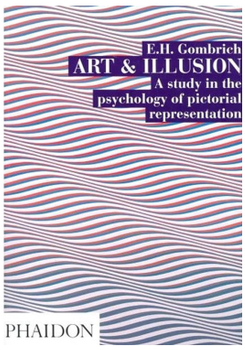


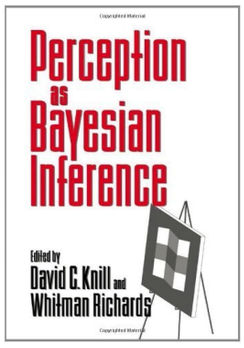
Ambiguity as to 3D Shape

Generally vision has been thought of as an Cognitive Inference or "guess" about the world since at least the 1960s with people like Richard Gregory and Irvin Rock.







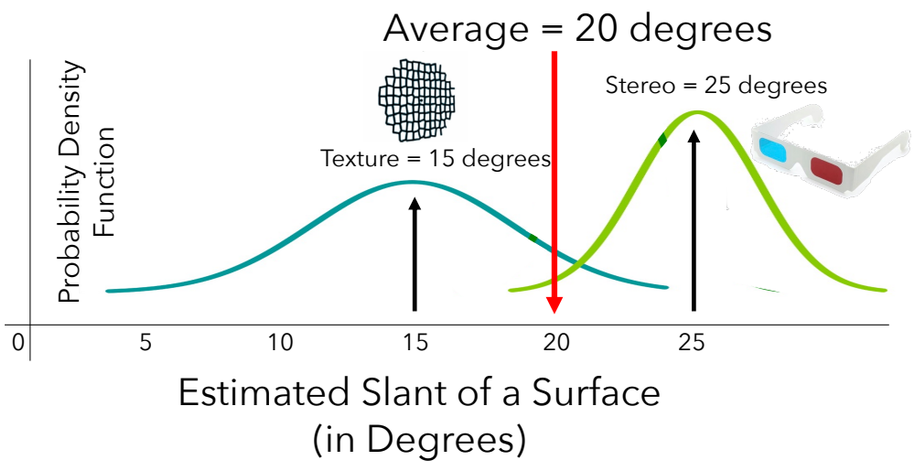


This way of thinking is also reflected in Gombrich's "beholder's share", **so I'm expecting serious insights from the art historians!** This way of thinking has continued to dominate for 60+ years. And the last 30 years has seen a particular version of it come to the fore: "Perception as Bayesian Inference". The basic idea of Bayesian inference is that you weight different interpretations of the sensory data that come in according to how likely they are given your previous experience.

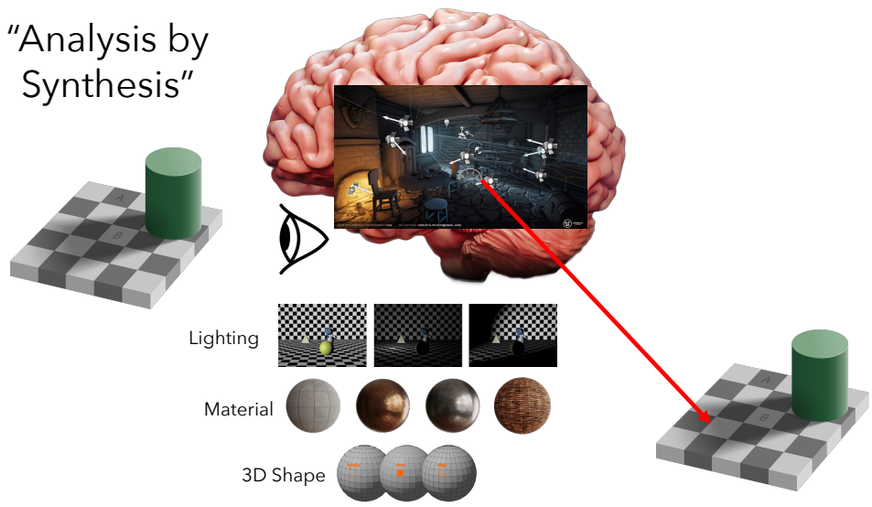
So, to give an example, what is the probability I'm seeing someone wearing a Manchester United shirt if I see someone wearing a red football shirt? Very high if I'm in Manchester, but relatively low I'm in London (more likely to be Arsenal).

Ok, so what does this look like as a model of vision? Well traditional accounts of Bayesian perception (in the Knill and Richards book, above) is something like this. Imagine you're trying to estimate the slant of a 3D surface. You might get an estimate from texture/perspective that says the slant is 15 degrees. And you might get an estimate from stereo vision saying that the slant is 15 degrees. Vision is supposed to take an average between the two (20 degrees). In practice, it takes a weighted average depending on how certain each cue is about its estimate. See how stereo

vision's estimate isn't that wide – it's pretty certain it's around 25 degrees – so you might weight your overall estimate more towards that, so average = 22 degrees.



More recent versions of 3D vision go even higher level still. So Josh Tenenbaum at MIT is famous for advocating an "inverse graphics" approach to human 3D vision.



On this approach, the human brain simulates various different versions of the scene in a "game engine" in the head – using different combinations of Lighting, Material, and 3D Shapes – until it gets a good match. And then, because the brain accurately simulated the scene, it has a good idea of what 3D Shape etc. caused it.

By contrast, I don't think that 3D vision is anything like this. It's not trying to work out the properties of the physical world by any kind of guesswork. Indeed, I would argue that it's not trying to work out the physical properties of the world at all!

Early 1600s

The high-level "Cognitive" account is something that's been historically dated back to Ibn al-Haytham's *Book of Optics* (c.1021). But my own research focuses on a much easier case to think about: How do we get 3D vision from two eyes?

Kepler and Descartes in the early 1600s had this well worked out with their **'triangulation'** theory of vision. Illustration from Descartes' *Treatise on Man* (1630s). So the account is very simple. Because we have two eyes now, instead of one, the inverse optics problem solves itself. Light travels in a straight line from points in the world (A, B, C) to points on the retina, and so all the visual system has to do is project

those points on the retina back into the world, and work out where they intersect, to recover the location of points in the world. And this has been the leading account of stereo vision for 400 years. There is a slight complication to this account, that Kepler and Descartes immediately noticed and incorporated. Unlike many animals, humans have rotating eyes, and the points projected on the retina depend on how the eyes are rotated. So instead of "points on retina → points in the world" you now need to go "points on retina → work out eye rotation → points in the world". But Kepler and Descartes presupposed that the human visual system simply knew where the eyes were pointing.

And so this is pretty much the account of stereo vision we have today. Going from points on the retina back out to points in the world.

QUIZ TIME!

Vision was a hot topic in the early 1600s, and everyone had treatises on Optics!

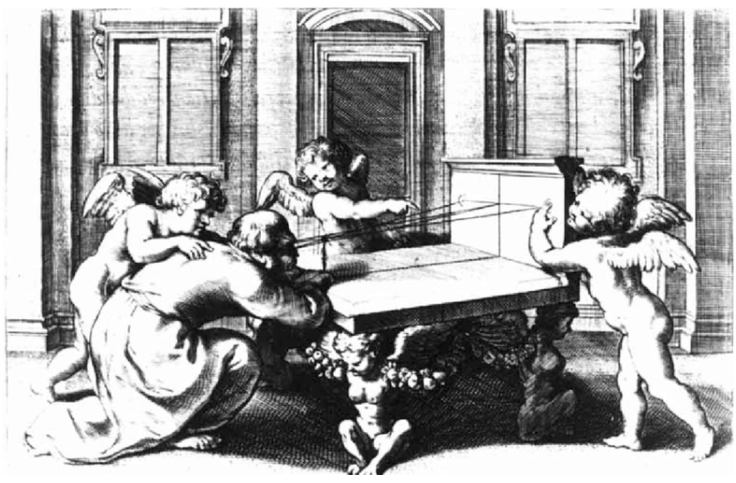
Question 1: (A) below is an illustration from François d'Aguilon's treatises on Optics (1613), but who was the artist who did the illustrations?

Question 2: (B) below is from someone else's treatises on Optics (1646, never published), this time someone remembered for political philosophy. Who was it?

Question 3: Who did the illustration in (B) below? Someone who later came to be regarded as one of the first economists / statisticians?

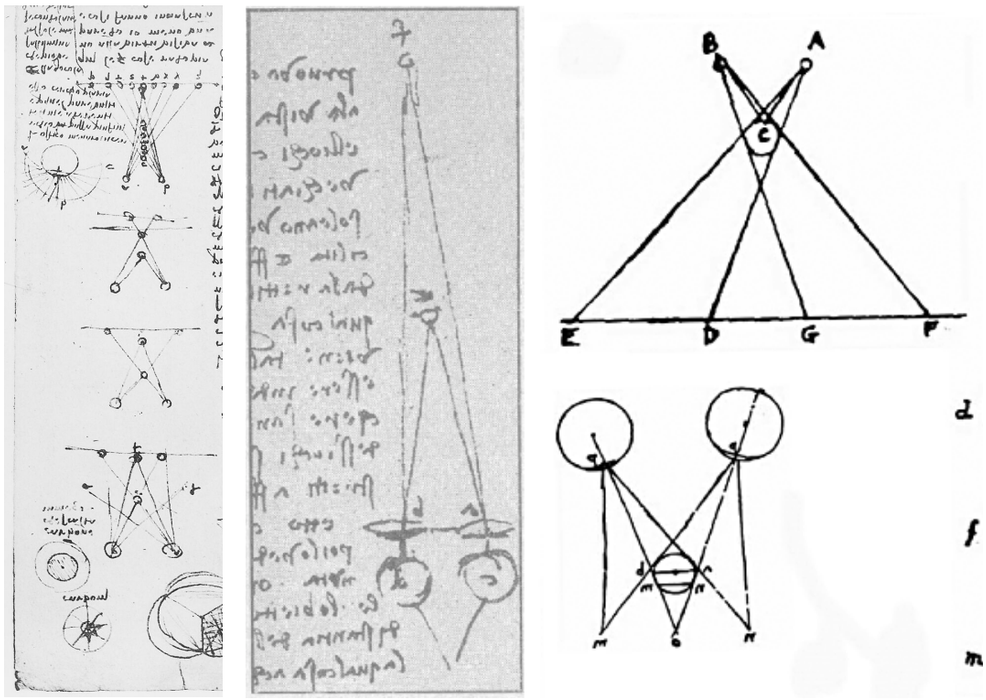
Question 4: What needed correcting in (B) below?

A



B

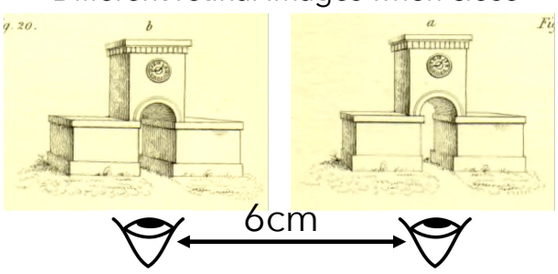
Question 5: Which 15th Century figure almost pre-empted the Descartes model?



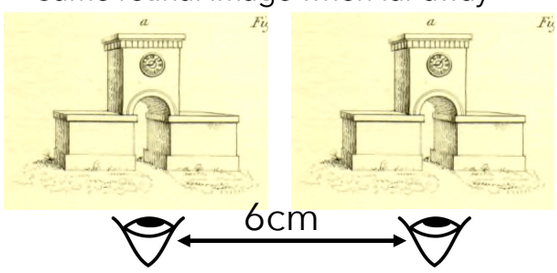
Early 1800s

There's two ways of thinking about vision. One is as an external observer of someone perceiving (bird's eye view of the process). Another is as the observer themselves. The big leap that Charles Wheatstone made in the early-1800s (1838 to be precise), was to notice how the top-down geometry described by Descartes manifested itself in the retinal image the observer sees. The simple point is that when an object is close, the spacing between the eyes leads to a different perspective in each eye:

Different retinal images when close

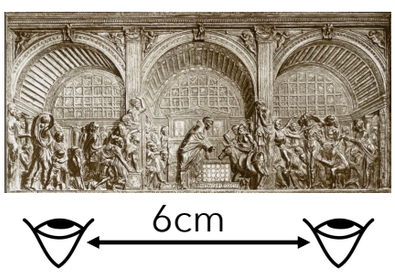


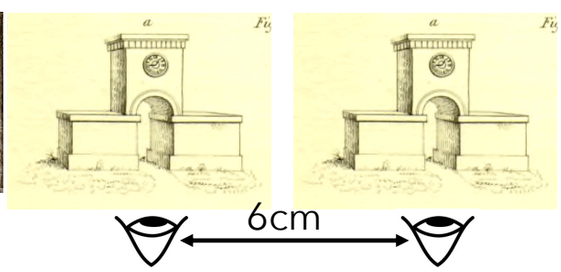
Same retinal image when far away



Now it's said that the only reason the person in **Question 5** missed this is because their object was a round ball (vs a square cube), so no change in perspective. But I don't think that's it. I think the key advance was to stop being a scientist – stop asking what it looks like from the "outside" – and start being an artist – start asking what it looks like from the "inside". If only the person in **Question 5** had been an artist!

But we have a problem. What happens when you look at something flat up close?





The projected retinal image of a flat object up close looks very similar to a 3D object viewed from far away! In both cases, the retinal images are of the two eyes are almost identical. So Wheatstone's "internal point of view" (vs Descartes' "external point of view") highlights an important dilemma – we need to know the distance of the object in order to disambiguate whether it is a flat object up close or a 3D object far away.

And I would argue you have 3 options:

1. **Simple Mechanism**

2. **Go Cognitive**

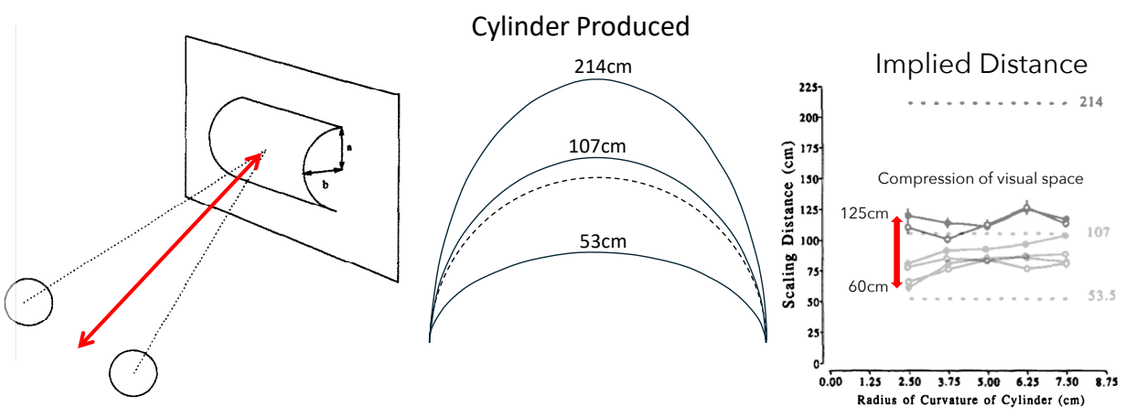
3. **Go Linton**

Simple Mechanism: Does the visual system have a distance range-finder? Some animals, like bats, have echolocation? And historically, 'extramission' theories of vision hypothesised that humans emitted light rays that returned to the eye. But neither of these two seem to be options. So how can human vision estimate distance?

Well, the closer an object is, the more the two eyes have to rotate inwards in order to fixate on it. This is known as **"vergence"**, and since Kepler and Descartes "vergence" has been thought of our most important source of absolute distance information, especially at close distances. **Questioning this is central to my project!**

"Vergence" is thought to provide a noisy but compressed source of distance information to enable the visual system to recover 3D shape in stereo vision.

The classic experiment of this is Johnston (1991). She had participants view a side on cylinder, and they had to set it so that it looked regular, i.e. 'b' = 'a' in the left picture (the two circles = observer's eyes). What she sound is that the shape that people set 'b' to depended on the viewing distance (red arrow in left picture). So at 107cm they were close to perfect (perfect = the dotted line). But at 53cm, the cylinder they produce is flattened, suggesting that stereo vision must be accentuated at close distances (because this looks like a regular cylinder to them). Conversely, at 214cm, the cylinder they produce is accentuated in depth, suggesting that stereo vision must be flattened at far distances (because this looks like a regular cylinder to observers).



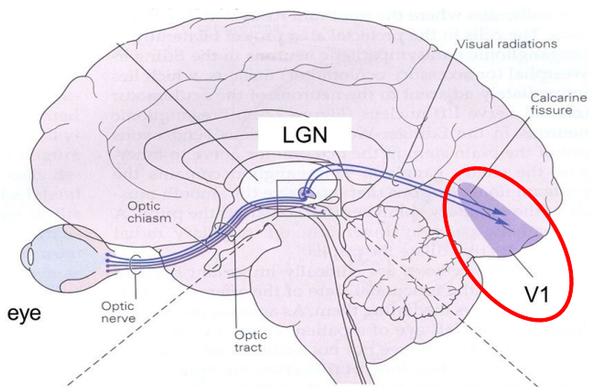
The explanation in the literature (right panel) is that the distance estimates from vergence are compressed. So 53cm is internally reported as 60cm. 214cm is internally reported as 125cm. But the general idea of using a distance estimate from vergence to work out 3D shape remains intact. **And that is what I want to challenge!**

Go Cognitive: You might think there's a bunch of other distance cues you can use to tell whether what you are looking at is near or far and use these to make sense of the projections on the retina. But the beginning of the talk will question this.

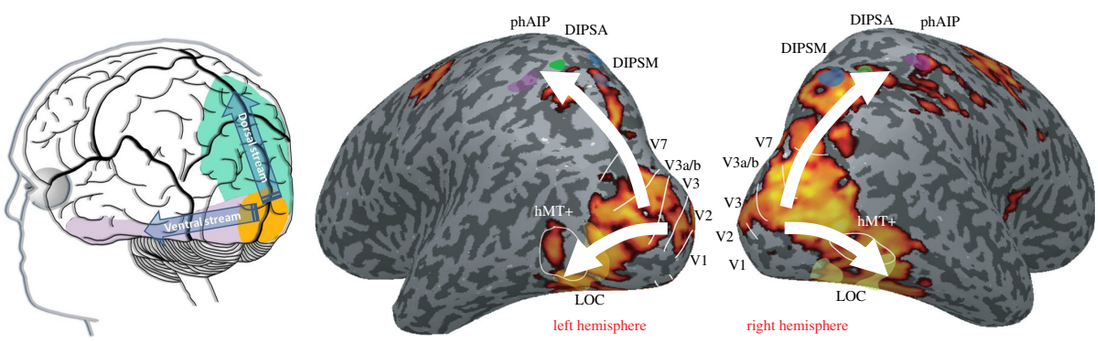
Go Linton: The alternative is just to accept that looking at a flat object up close and a 3D object far away gives the very same stereo vision experience. There is 60 years' worth of experimental data claiming to show that it does not. But in my talk I will discuss a new experimental paradigm (the 'Linton Stereo Illusion') that challenges this. I have attached to the abstract describing this to the end of this pre-paper.

The Brain

Finally, we want our theories about Cognitive Science (how 3D vision works) to tell us something about Neuroscience (where in the brain stereo vision, and 3D vision in general, is processed). My bet would be the Primary Visual Cortex (V1):



So light hits the retina, then the signal travels to LGN, then travels to V1 (Primary Visual Cortex). Simple, right? Well, not so much. You see the 3D vision literature have come to the following conclusion. If there's one thing we know about 3D vision, it is that it is not processed in the Primary Visual Cortex (V1), but much further along the Dorsal (top of the brain) and Ventral (bottom of brain) Streams, indicated by white arrows.



So see how in these brain activation maps V1 is <u>not</u> "lit up" (it's dark), but the Dorsal and Ventral Streams are lit up. **But I argue they are looking for the wrong thing!**

Perceived Stereo Depth reflects Retinal Disparities, not 3D Geometry

Paul Linton[1,2,3] & Nikolaus Kriegeskorte[3,4,5,6]

[1] Presidential Scholars in Society and Neuroscience, Center for Science and Society, Columbia University
[2] Italian Academy for Advanced Studies in America, Columbia University
[3] Visual Inference Lab, Zuckerman Mind Brain Behavior Institute, Columbia University
[4] Department of Psychology, Columbia University
[5] Department of Neuroscience, Columbia University
[6] Department of Electrical Engineering, Columbia University

Keywords: Stereo Vision, Depth Constancy, 3D Vision

We present a new illusion that challenges our traditional understanding of stereo vision. Traditional 'Triangulation' accounts of stereo vision back-project from points on the retina to points in the world. This requires that stereo vision incorporates how binocular disparities fall off with the viewing distance squared. By contrast, Linton 2023 *Phil Trans R Soc B* 378: 20210455 proposes a 'Minimal Model' of stereo vision where perceived stereo depth is simply a function (most likely a linear function) of the amount of disparity on the retina. We present a new illusion (the 'Linton Stereo Illusion') to adjudicate between these two approaches. The illusion consists of a smaller circle (at 40cm) in front of a larger circle (at 50cm), with constant angular sizes throughout. We move the larger circle forward by 10cm (to 40cm) and then back again (to 50cm). The question is, what distance should we move the smaller circle forward and back to maintain a constant perceived separation between the circles? Constant physical distance (10cm) ('Triangulation') or constant disparity (6.7cm) ('Minimal Model')? Observers choose constant disparity. This leads us to four conclusions: First, perceived stereo depth appears to be best captured by the 'Minimal Model'. Second, doubling disparity appears to double perceived depth, suggesting that perceived stereo depth is proportional to disparity. Third, changes in vergence appear to have no effect on perceived depth. Fourth, stereo 'depth constancy' appears to be a cognitive (not perceptual) phenomenon, reflecting our experience of a world distorted in perceived stereo depth.

Funding: Presidential Scholars in Society and Neuroscience, Center for Science and Society, Columbia University + Italian Academy for Advanced Studies in America, Columbia University

Abstract Presented at Applied Vision Association Spring Meeting 2024