PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Introduction

Cite this article: Linton P, Morgan MJ, Read JCA, Vishwanath D, Creem-Regehr SH, Domini F. 2022 New Approaches to 3D Vision. *Phil. Trans. R. Soc. B* **378**: 20210443.
https://doi.org/10.1098/rstb.2021.0443

Received: 30 June 2022
Accepted: 25 October 2022

One contribution of 18 to a discussion meeting issue 'New approaches to 3D vision'.

Subject Areas:
artificial intelligence, vision, robotics, cognition, neuroscience

Keywords:
3D vision, artificial intelligence, computer vision, navigation, human vision

Author for correspondence:
Paul Linton
e-mail: paul@linton.vision

THE ROYAL SOCIETY PUBLISHING

New Approaches to 3D Vision

Paul Linton[1,2,3], Michael J. Morgan[4], Jenny C. A. Read[5], Dhanraj Vishwanath[6], Sarah H. Creem-Regehr[7] and Fulvio Domini[8]

[1]Presidential Scholars in Society and Neuroscience, Center for Science and Society, Columbia University, New York, NY 10027, USA
[2]Italian Academy for Advanced Studies in America, Columbia University, New York, NY 10027, USA
[3]Visual Inference Lab, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA
[4]Department of Optometry and Visual Sciences, City, University of London, Northampton Square, London EC1V 0HB, UK
[5]Biosciences Institute, Newcastle University, Newcastle upon Tyne, Tyne & Wear NE2 4HH, UK
[6]School of Psychology and Neuroscience, University of St Andrews, St Andrews, Fife KY16 9JP, UK
[7]Department of Psychology, University of Utah, Salt Lake City, UT 84112, USA
[8]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI 02912-9067, USA

PL, 0000-0002-6964-9575; MJM, 0000-0002-5946-3101; JCAR, 0000-0002-9029-5185; DV, 0000-0002-8095-0537; SHC-R, 0000-0001-7740-1118; FD, 0000-0002-5510-0397

New approaches to 3D vision are enabling new advances in artificial intelligence and autonomous vehicles, a better understanding of how animals navigate the 3D world, and new insights into human perception in virtual and augmented reality. Whilst traditional approaches to 3D vision in computer vision (SLAM: simultaneous localization and mapping), animal navigation (cognitive maps), and human vision (optimal cue integration) start from the assumption that the aim of 3D vision is to provide an accurate 3D model of the world, the new approaches to 3D vision explored in this issue challenge this assumption. Instead, they investigate the possibility that computer vision, animal navigation, and human vision can rely on partial or distorted models or no model at all. This issue also highlights the implications for artificial intelligence, autonomous vehicles, human perception in virtual and augmented reality, and the treatment of visual disorders, all of which are explored by individual articles.

This article is part of a discussion meeting issue 'New approaches to 3D vision'.

In November 2021 we held a Royal Society scientific meeting on 'New approaches to 3D vision' with the following mission statement:

> Leading approaches to computer vision (SLAM: simultaneous localization and mapping), animal navigation (cognitive maps), and human vision (optimal cue integration), start from the assumption that the aim of 3D vision is to produce a metric reconstruction of the environment. Recent advances in machine learning, single-cell recording in animals, virtual reality, and visuomotor control, all challenge this assumption. The purpose of this meeting is to bring these different disciplines together to formulate an alternative approach to 3D vision.

And now was the perfect time to host this meeting. With artificial intelligence's success in 2D vision, attention is now turning to 3D vision. There's been an explosion of interest in 3D image reconstruction ('A New Trick Lets Artificial Intelligence See in 3D', *Wired Magazine* [1]), considerable successes in using 3D vision to uncover new biological advances (with DeepMind's AlphaFold [2,3] solving the protein-folding problem), and the suggestion that grounding artificial intelligence in 3D vision will enable better AI (MURI[1], [4–6]). But 3D vision still remains a challenge for AI [7], and is often regarded as the most difficult question facing robotics and autonomous vehicles [8–10].

At the same time, we are also seeing considerable advances in our understanding of biological vision and navigation. Single-cell recording in freely moving animals has enabled us to understand for the first time how the brain's map of 3D space is organized [11,12], while the emergence of virtual and augmented reality has

Table 1. Schedule of talks for the Royal Society meeting 'New Approaches to 3D Vision' (1–4 Nov 2021). For links see: https://osf.io/2waby.

NEW APPROACHES TO 3D VISION
Royal Society, 1-4 Nov 2021
Website / Recordings

DAY ONE (1st Nov) – Seeing Beyond SLAM
Chair: Andrew Fitzgibbon (Microsoft)

Session One: **Neural Scene Representation**
SM Ali Eslami (DeepMind): "Neural priors, neural encoders and neural renderers"
Ida Momennejad (Microsoft Research): "Multi-scale predictive representations and human-like RL"
Session One Discussion (Fitzgibbon / Eslami / Momennejad)

Session Two: **Perception-Action Loop**
Sergey Levine (UC Berkeley and Google): "Generalization in data-driven control"
Andrew Glennerster (University of Reading): "Understanding 3D vision as a policy network"
Session Two Discussion (Fitzgibbon / Levine / Glennerster)

DAY TWO (2nd Nov) – Animals in Action
Chair: Matteo Carandini (University College London)

Session One: **Locating Prey and Rewards**
Jenny Read (Newcastle University): "Stupid stereoscopic algorithms that still work"
Aman Saleem (University College London): "Visual processing in the brain during navigation"
Session One Discussion (Carandini / Read / Saleem)

Session Two: **Navigation in 3D Space**
Kate Jeffery (University College London): "The cognitive map of 3D space: not as metric as we thought?"
Gily Ginosar (Weizmann Institute of Science): "Locally ordered representation of 3D space in the entorhinal cortex"
Session Two Discussion (Carandini / Jeffery / Ginosar)

DAY THREE (3rd Nov) – Experiencing Space
Chair: Mar Gonzalez-Franco (Microsoft Research)

Session One: **Theories of Visual Space**
Dhanraj Vishwanath (University of St Andrews): "Tripartite encoding of visual 3D space"
Paul Linton (City, University of London): "New approaches to visual scale and visual shape"
Session One Discussion (Gonzalez-Franco / Vishwanath / Linton)

Session Two: **Challenges for Virtual Reality**
Sarah Creem-Regehr (University of Utah): "Perception and action in virtual and augmented reality"
Douglas Lanman (Facebook Reality Labs): "Engineering challenges for realistic displays"
Session Two Discussion (Gonzalez-Franco / Creem-Regehr / Lanman)

DAY FOUR (4th Nov) – Grasping the World
Chair: Jody Culham (Western University)

Session One: **One Visual Stream or Two?**
Fulvio Domini (Brown University): "A novel non-probabilistic model of 3D cue integration explains both perception and action"
Irene Sperandio (University of Trento): "Dissociations between perception and action in size-distance scaling"
Session One Discussion (Culham / Domini / Sperandio)

Session Two: **3D Space and Visual Impairment**
Ione Fine (University of Washington): "Do you hear what I see? How do early blind individuals experience object motion?"
Ewa Niechwiej-Szwedo (University of Waterloo): "The role of binocular vision in the development of visuomotor control and performance of fine motor skills"
Session Two Discussion (Culham / Fine / Niechwiej-Szwedo)

Session Three: **Future Directions**
Chair: Michael Morgan FRS (City, University of London)
Panel Discussion by the Chairs (Fitzgibbon / Carandini / Gonzalez-Franco / Culham)

required that we reconsider the fundamental principles underpinning human 3D vision.

Over 800 people participated in our meeting, with speakers from DeepMind, Google Robotics, Microsoft Research, and Meta (Facebook) Reality Labs, as well as academics from both basic and applied research. Recordings and abstracts of the talks are available on the Royal Society website [13], and links to the talks available in table 1.

The purpose of our meeting was to capitalize on a brief moment when computer vision, animal navigation, and human vision are all pausing and asking what the most appropriate representation for 3D vision and action really is? On the one hand, it's natural to think that the purpose of 3D vision is to provide us with an accurate model of the environment. On the other hand, in recent years computer vision, animal navigation, and human vision have all been grappling with whether a partial, distorted, or even inconsistent model of the environment might suffice, or perhaps no model at all. And the hope is that by drawing attention to the similarity of these discussions in computer vision,

animal navigation, and human vision, we can begin to connect these different approaches, which have evolved relatively independently of one another.

In this Introduction, we outline how these different disciplines have tackled this question and highlight the specific contributions that the papers in this issue make.

1. Computer vision

Computer vision originated with 3D vision in the 1960s (Larry Roberts' 'Machine Perception of Three Dimensional Solids' (1963) [14–16]). And the emphasis in the 1960s–80s was on using computer vision to reconstruct an accurate 3D model of the environment:

> …vision is the *process* of discovering from images what is present in the world, and where it is. (Marr, [17])

The dominant approach to solving this problem was 'analysis by synthesis', which treats vision as 'inverse optics' or 'inverse graphics'. For instance, Waltz [18]:

> The overall goal of the system is to provide a precise description of a plausible scene which could give rise to a particular [image]…

Horn [19]:

> The problem can be viewed as one of inversion: If we understand the projection process which creates images from the three dimensional world, we can hope to reverse this process to recover information about the world.

And Charniak & McDermott [20]:

> Given a 2D image, infer the objects that produced it, including their shapes, positions, colors, and sizes.

However, by the early 1980s the pace of progress on 3D world models had slowed ([21], p.ix). And in the mid 1980s and early 1990s, the necessity of an accurate 3D world model began to be questioned on two fronts: Active Vision and Non-Euclidean Geometries.

(a) Active vision

Advances in computer hardware meant real-time active robotics became a possibility in the mid-1980s. Ruzena Bajcsy at Penn [22,23] and Chris Brown and Dana Ballard at Rochester [24–26] developed robots with active eye movements, complementing previous work by Marty Tenenbaum [27] on computer vision using active lens focusing.

In the mid-1980s, Rodney Brooks [28] developed state-of-the-art autonomous robots by explicitly rejecting the need for a 3D world model [29]:

> Internal world models which are complete representations of the external environment, besides being impossible to obtain, are not at all necessary for agents to act in a competent manner

And this was a feature of many Active Vision discussions:

> …vision is more readily understood in the context of the visual behaviors that the system is engaged in, and that these behaviors may not require elaborate categorical representations of the 3-D world. (Ballard [25])

> The notion of direct coupling of perception and action, without an explicit 3D intermediary, is very appealing. (Blake & Yuille [30], p.173)

Two volumes [30,31] summarize the state of the art up to the early 1990s. And [26] highlights three ways in which Active Vision 'recast completely the role of vision', which was adhered to (to varying degrees) by different advocates of Active Vision.

First, active vision is task-specific: 'an active vision system is far more selfish. It picks out the properties of images which it needs to perform its assigned task, and ignores the rest.' (Blake & Yuille, [30], p.xv). This was an insight from Yarbus [32]'s work on human eye movements, where patterns of eye movements changed depending on the task.

Second, active vision is dynamic. It only extracts what it needs now. It uses the physical world as its own best model, sampling the physical world when and where it needs to: 'the visual scene acts as a kind of *external memory buffer* whose unclear parts can be activated by making an eye movement' (O'Regan & Lévy-Schoen [33]). On this account, 'the world is its own best model. … The trick is to sense it appropriately and often enough.' (Brooks [34]).

Third, active vision is adaptive. Since action dictates what visual information is picked-up from the environment, vision is shaped by the organism's interactions with the world, and must be responsive to it. This eradicates a sharp distinction between 'perception' and 'motor control' modules: 'there need be no clear distinction between a 'perception subsystem', a 'central system' and an 'action system' (Brooks' [35]). Instead, on Active Vision accounts, modules are tied to specific tasks and behaviours: 'In the purest form of this model each module incorporates its own perceptual, modelling and planning requirements'. (Brooks [36]).

(b) Non-Euclidean geometries

In the early 1990s there was also an explosion of interest in using non-Euclidean geometries to solve problems in computer vision:

> We usually think of physical space as being embedded in a 3D Euclidean space, in which measurements of length and angles do make sense. It turns out that for artificial systems, such as robots, this is not a mandatory viewpoint and that it is sometimes sufficient to think of physical space as being embedded in an affine or even a projective space. (Faugeras [37])

Affine geometry captures geometry in a loose sense, preserving parallel lines, but not distances or angles. Projective geometry fails to preserve even parallel lines.

First, it was realised that many tasks can be accomplished without a metric model of the environment. There was a sense that 'computer vision may have been slightly overdoing it in trying at all costs to obtain metric distance information from images.' (Faugeras [38]). By contrast, 'Affine structure offers a useful compromise between difficulty of computation and information content.' (Beardsley *et al*. [39]). Consequently, affine accounts of structure from motion [40], stereo vision [38,41], navigation [39,42,43], and object recognition [44,45] soon emerged.

Second, even when metric scene estimates are necessary, they can be achieved more easily and directly simply by adding constraints to affine geometry, rather than attempting full 3D scene reconstruction: 'one can estimate all 3D invariants of the scene directly from the images, *without* performing an explicit 3D reconstruction of the scene.' (Faugeras [37]).

(c) LIDAR

By contrast, the early 2000s saw a strong re-emergence of Euclidean 3D maps in computer vision. Two developments, LIDAR and SLAM, were key to this re-emergence.

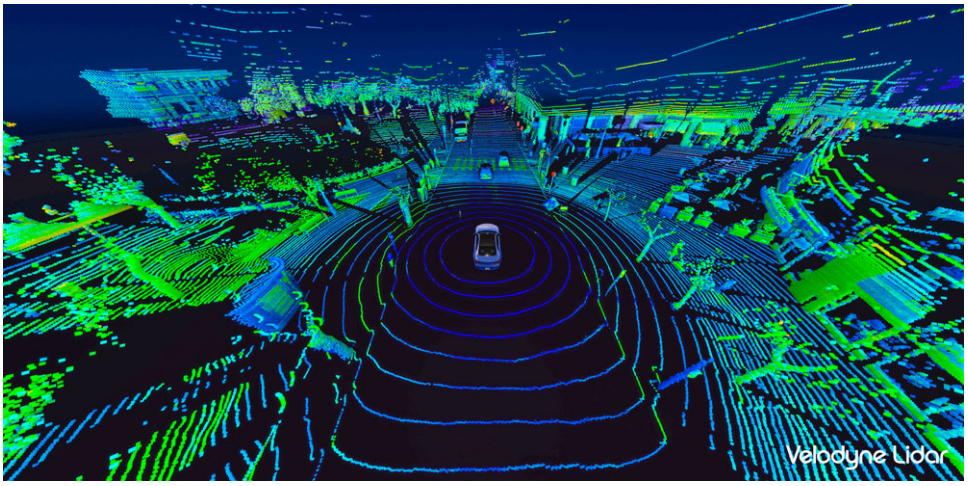

Figure 1. The 360° point cloud created by the Alpha Prime Velodyne Lidar. © Velodyne Lidar. (Online version in colour.)

LIDAR (light detection and ranging) was invented in the 1960s [46]. It estimates the distance of an object in a certain direction by emitting a laser pulse and timing how long it takes to return. LIDAR became synonymous with computer vision in 2005 when 'Stanley' (a LIDAR-equipped car) won the 2nd self-driving car DARPA Grand Challenge [47] (previously, all cars had failed the 1st DARPA Grand Challenge). Velodyne's 360° LIDAR was also launched during the 2nd DARPA Grand Challenge, producing a high resolution 360° map of the environment (figure 1). This soon became the industry standard, used by 5 of the 6 finishers of the 3rd DARPA Grand Challenge in 2007 [48,49] as well as Google's self-driving car project (Waymo) [49].

However, the success of deep learning since 2012 [50] has challenged the importance of LIDAR for self-driving cars, leading to a split in the industry. While LIDAR remains a key feature of most self-driving car projects (Waymo (Google), Baidu, Cruise (General Motors)), LIDAR has been explicitly rejected by Tesla since 2013, when its self-driving car program began [51]. Indeed, in 2021 Tesla removed the only range-finding component (radar) from their self-driving cars, with a rationale that explicitly evokes human vision: 'Humans drive with eyes & biological neural nets, so makes sense that cameras & silicon neural nets are only way to achieve generalized solution to self-driving.' [52].

(d) SLAM

SLAM (simultaneous localization and mapping) is the ability of a robot to build a map of its environment, whilst locating itself relative to this map as it navigates. As Thrun *et al.* [53] noted in 2000:

> Building maps when a robot's locations are known is relatively straight-forward … Conversely, localizing a robot when a map is readily available is also relatively well understood … In combination, however, the problem is hard.

SLAM remained an intractable problem until the early 2000s. As Durrant-Whyte and colleagues [54] wrote in 1996:

> The solution to the simultaneous localisation and map building (SLAM) problem is, in many respects a 'Holy Grail' of autonomous vehicle navigation research.

But by 2006, Durrant-Whyte & Bailey [55] could reasonably claim that:

> The 'solution' of the SLAM problem has been one of the notable successes of the robotics community over the past decade.

So, what changed between 1996 and 2006?

First, the emergence of 'Probabilistic Robotics' [56] led to a number of Bayesian solutions to the SLAM problem. The first, and by far the most influential, was Extended Kalman Filter (EKF) SLAM (e.g. [57]), derived from Smith *et al.* [58]'s concept of a 'stochastic map':

> …rather than treat spatial uncertainty as a side issue in geometrical reasoning, we believe it must be treated as an intrinsic part of spatial representations. In this paper, spatial uncertainty will be tied together in a representation called the *stochastic map*. It contains estimates of all the spatial relationships, their uncertainties, and their inter-dependencies.

Another notable solution was FastSLAM [59,60].

Second, early attempts at SLAM combined odometry (motion sensors) with range-finders like sonar [61], radar [62], or lidar [63]. In the early 2000s vision-based systems (Visual SLAM) began to supersede these approaches. Davison & Murray combined SLAM with Active Vision to enable autonomous navigation for a robot with an active stereo head [64–66]. And with MonoSLAM, Davison and colleagues went one step further, using structure from motion to enable SLAM with a single moving camera with no motion sensors or motor commands [67,68].

However, SLAM has two notable shortcomings. First, SLAM is unable to build-up an intuitive understanding of the environment (Gupta *et al.* [69]): 'These maps are built purely geometrically, and nothing is known until it has been explicitly observed, even when there are obvious patterns.' New approaches therefore seek to augment SLAM with deep learning [70–74] (see also 3D semantic scene graphs [75–77]). Others seek an alternative to SLAM in deep reinforcement learning [69,78–83] or deep learning [84].

Second, and more fundamentally, SLAM is biologically implausible:

> …humans can effectively navigate small and large environments and yet are unlikely to build internally large-scale metric reconstructions of spaces akin to traditional SLAM systems… (Henriques & Vedaldi [84])

> Look at the inner workings of most map building algorithms … and there is a strong likelihood of finding Cartesian (x, y, z) representations of the locations of features. It is not clear this is the best way to do things on such a low level, or that biological brains have any similar representation… (Davison [66], p.1)

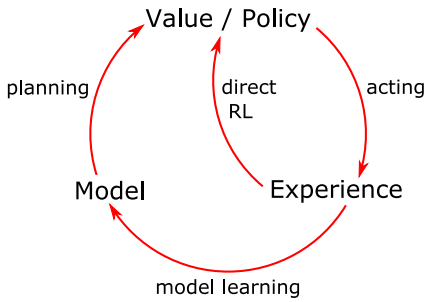

Figure 2. Diagram of Dyna-Q [85] redrawn from [86], p.7, which incorporates both 'model-based' ('value/policy' → 'experience' → 'model' → 'value/policy') and 'model-free' ('value/policy' → 'experience' → 'value/policy') reinforcement learning. (Online version in colour.)

Especially when SLAM is given an explicit Extended Kalman Filter articulation:

> It certainly does not seem that we store the equivalent of a huge covariance matrix relating to the uncertainties in our estimates of the positions of all the features in a certain area. (Davison [66])

This suggests that the extensive metric reconstruction of the environment proposed by SLAM is not necessary for vision or navigation, and is therefore unnecessarily complicated.

(e) Reinforcement learning

An alternative approach is reinforcement learning. Reinforcement learning is the process of learning which action to take to maximize the rewards in a given context. Reinforcement learning can take two forms: model-based and model-free, as summarized in figure 2.

Model-based reinforcement learning learns a model of the world in order to predict the rewards of potential actions. By contrast, model-free reinforcement learning learns a direct mapping from the inputs to the actions that maximize rewards.

The earliest work on reinforcement learning was model-based. Richard Sutton and Andrew Barto saw themselves as building 'an adaptive network that constructs and uses an internal model of its world' [87], as envisioned by Kenneth Craik:

> If the organism carries a 'small scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best for them, react to future situations before they arise… ([88] quoted in [87])

By contrast, in the late 1980s 'model-free' approaches (such as 'temporal difference' [89] and 'Q-learning' [90,91]), which do away with a world model altogether, came to the fore. Model-free approaches are:

> explicitly trial-and-error learners – viewed as almost the *opposite* of planning. (Sutton & Barto [86], p.7)

And Sutton & Barto [86], p.12 explain the advantages of this approach:

> Because models have to be reasonably accurate to be useful, model-free methods can have advantages over more complex methods when the real bottleneck in solving a problem is the difficulty of constructing a sufficiently accurate environment model.

Reinforcement learning came to the public consciousness in 2015 when DeepMind's Deep Q-Network (DQN) achieved human-level performance in Atari computer games [92], and in 2016 when their AlphaGo defeated the world's top Go player [93]. Both relied on 'deep' model-free reinforcement learning, where the mapping from inputs to actions is learned by a deep (multilayer) neural network. (Earlier notable success, such as IBM's TD-Gammon [94], which performed competitively at the 1992 World Cup of Backgammon, relied on a 'shallow' (single layer) neural network.)

Model-free reinforcement learning challenges the traditional relationship between perception, planning, and action. When DeepMind's Deep Q-Network 'plays' Space Invaders, all it learns is the mapping between the pixels on the screen (input) and the buttons to press (output). This is why it's described as a 'pixels to action' approach [95]. And end-to-end pixel-to-action training has been successfully applied to robot object manipulation [96], navigation [97–100], driving [101,102], and 3D video games [103].

At our meeting, Sergey Levine [104] argued that this pixel-to-action approach improves performance since robots learn to extract the most relevant representations for the task from the visual input. Similarly, Zhou *et al.* [105] note in 'Does computer vision matter for action?'

> These models bypass explicit computer vision entirely. They do not incorporate modules that perform recognition, depth estimation, optical flow, or other explicit vision tasks. The underlying assumption is that perceptual capabilities will arise in the model as needed, as a result of training for specific motor tasks. This is a compelling hypothesis that, if taken at face value, appears to obsolete much computer vision research.

But the rise of this pixel-to-action approach raises two fundamental questions.

First, would visuomotor control benefit from having an explicit 3D depth map? On the one hand, [105] found that having an explicit depth map and scene segmentation for input significantly improved performance in certain tasks and, even if it didn't (e.g. urban driving), it helped the model to generalize to new and unseen environments (see also [106–111]). And [69,78] also argue that navigation is improved if agents are able to build a top-down Euclidean map of the scene on which to plan their strategies. On the other hand, in this issue Levine & Shah [104] argue that what matters for navigation is traversability, not 3D geometry: tall grass is traversable, even though it looks like a barrier, whilst mud is not traversable, even though it looks like a flat surface, and an explicit representation of 3D geometry is an unnecessary bottleneck to learning traversability. Finally, a third alternative is to use a depth map, but to learn the depth map as part of reinforcement learning, rather than as an input to reinforcement learning [98].

Second, even if this pixel-to-action approach doesn't have an explicit 3D map of the environment, does it effectively learn an implicit 3D map? As Zhu *et al.* [97] note: 'Our method is considered *map-less*. However, it possesses implicit knowledge of the environment.' What is the nature of this implicit knowledge? This question has sparked collaborations between computer vision and psychology. On the one hand, the Glennerster & Torr labs study the specific spatial representation in Zhu *et al.* [97], and find 'only a weak correlation between distance in the embedding space and physical distance between observable locations' [112]. By contrast, SoftBank Robotics and Kevin O'Regan explore the possibility that a Euclidean representation of space emerges organically from reinforcement learning [113–116].

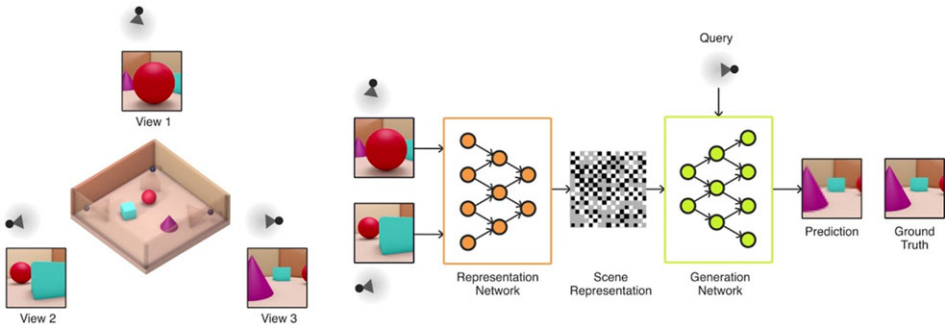

Figure 3. DeepMind's 'neural scene representation and rendering' [106]. First, the 'Representation Network' creates a low-dimensional 'Scene Representation' based on images from a number of different views (View 1 and View 2). Second, the 'Generation Network' uses this low-dimensional 'Scene Representation' to predict what View 3 will be. © authors. (Online version in colour.)

(f) Deep learning

The Summer Vision Project (1966) [117] defined the mission statement for computer vision as moving from 2D pixels → 3D surfaces → object recognition. But when deep neural networks finally achieved human levels of performance in object recognition, they did this by skipping the 3D surfaces step and going straight from 2D pixels → object recognition [50]. Contrast this with the Summer Vision Project (1966) [117] which assumed: 'It will be impossible to do this without considerable analysis of shape and surface properties'.

Indeed, 3D vision appeared something of a sticking point for neural networks. Henriques & Vedaldi [84] summarized the position in 2018:

> Despite these successes [in 2D images], … several aspects of image understanding remain difficult to approach directly using deep distributed representations. One of them is reasoning about 3D space and geometry…

So, can we teach deep neural networks to reason about 3D space? Here we have to be careful to distinguish neural networks that reason about 3D space from neural networks *that merely act as if they do*. Two general approaches have emerged:

1. Geometric deep learning (Explicit 3D model): Neural networks can be trained to do traditional 3D computer vision. For instance, take a single 2D image as an input and output an explicit 3D model of the object, either as voxels [118–120], point clouds [121], or meshes [122,123]. But it's unclear whether these neural networks are actually reasoning in 3D space (performing a spatial reconstruction of the scene) or merely performing object recognition (searching for a matching 3D template). In 'What Do Single-view 3D Reconstruction Networks Learn' [124], three of the then leading models [125–127] are tested, with the conclusion that the 'current state of the art in single-view object reconstruction does not actually perform reconstruction but image classification.' Indeed, for many the blurring of 3D reconstruction and object recognition is intentional [118–120,128–130].

2. Neural scene representations (Implicit 3D model): A different approach is to have neural networks learn 3D spatial layout implicitly rather than explicitly. At our meeting, Ali Eslami presented DeepMind's landmark work on 'neural scene representation' [106,131] (figure 3). Their network is provided with 2D images of a scene, but is never explicitly told to build a 3D model. Instead, DeepMind reason that if their network can take 2D images, encode a low-dimensional description of the scene, and then use this low-dimensional description to imagine what the scene would look like from a new viewpoint, then the network must have implicitly learnt the 3D layout of the scene. Contrast this with models that just try to predict how 2D pixels on a screen will change with a specific action, without trying to build an intermediate 3D representation [132].

However, since the 'scene representation' (low-dimensional description) is not interpretable by humans (it's not an explicit 3D model), we have to do experiments to indirectly reveal the nature of the representation. First, Eslami *et al.* [106] show that the same representation can be used to produce a top-down map of the scene, supporting the idea of an implicit 3D representation. However, they also note that changing the 2D input view of the scene changes the representation, even though it shouldn't if the network were learning a truly view-invariant representation of 3D space. Second, the tension between reconstruction and recognition re-emerges, with Tung *et al.* [73] arguing that failure cases in Eslami *et al.* [106] demonstrate that their 'geometry-unaware models may be merely memorizing views with small interpolation capabilities, as opposed to learning to spatially reason.'

Both [73] and [133] use failure cases in Eslami *et al.* [106] to argue that learning in explicit 3D coordinates is necessary for true 3D understanding. Consequently, the current literature is primarily focused on 'neural fields' [134], 'neural scene representations' that learn in explicit 3D coordinates.

What's interesting is that these models are implicit in a different sense. The scene representation isn't an output (an interpretable or non-interpretable model) that the network produces, but the neural network itself. So you input a point in 3D space, and the network outputs some property of that point in 3D space, for instance occupancy [135], distance [136], colour + distance [133] or colour + occupancy [137]. From this, the network implicitly learns a continuous (and uninterpretable) mathematical formula that approximates the scene.

Interest in these approaches exploded in 2020 with 'Neural Radiance Fields' (NeRFs) [138] (figure 4), a collaboration between Google, Berkeley and UC San Diego, which takes as inputs the location of a point in 3D space and the direction it's being viewed, and outputs the colour and density of the point in 3D space. This means it can capture the reflective properties of the surfaces, producing photorealistic 3D renderings of real scenes.

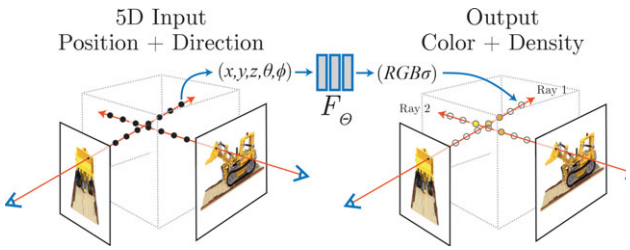

Figure 4. 'Neural Radiance Fields' (NeRFs) [138]. NeRFs generate images by sampling points along a ray (5D input = position + direction) and the network (F_Θ) outputs a colour and density value for each sampled point. NeRF is 'geometry aware' since its inputs are explicitly in 3D coordinates (x,y,z). © authors. (Online version in colour.)

However, two challenges remain:

First, like SLAM, 'Neural Radiance Fields' (NeRFs) are purely geometric. They can't use prior knowledge about the common structure of scenes to make common sense inferences. Instead, they rely purely on multi-view consistency, which can be achieved without deep learning [139]. This means the original NeRF paper required around 20–60 images per scene [138]. By contrast, recent deep learning approaches such as 'pixelNeRF' [140] and 'NeRF-VAE' [141] learn common image features across scenes, enabling new scenes to be reconstructed with just 2–3 images.

Second, NeRFs are biologically implausible. They work by giving a colour and density to each point in 3D space. And, to render a new image, they have to 'ray march' (by integrating the colours and densities of each point in 3D space that lie along the ray corresponding to each pixel (figure 4)). But this can't be what the human brain is doing. Instead, Vincent Sitzmann and Josh Tenenbaum [1] advance their 'light field network' [142] as a more biologically plausible approach because it estimates the colour of each ray (rather than each point along a ray), and builds on earlier work in human vision that employs light-field concepts ('plenoptic function' [143], 'optic array' [144]).

2. Animal navigation

Studying animals with very different eyes and brains from our own opens up new ways of thinking about vision and navigation.

First, the different optics of animal eyes may lead to different depth cues being prioritised. For instance, jumping spiders [145,146], squid [147], and even the prehistoric trilobite [148] have been shown to rely heavily on defocus blur [149], whilst chameleons appear to rely on accommodation (the ability to change the focus of the eye) [150].

Second, if animals have overlapping eyes, how they compute depth from binocular disparity (the difference in the images projected to the two eyes) may be very different [151–153]. In this issue we contrast humans and insects. Michael Morgan [154] explores the complexity of disparity processing in humans. By contrast, Jenny Read [155] explores how an animal with more limited computational capacities, such as the praying mantis, could extract distance information from disparity without the complexities of human vision (such as matching the points in the two eyes or extracting a depth map). Other studies of stereo vision in animals include owls [156], toads [157], and cuttlefish [158].

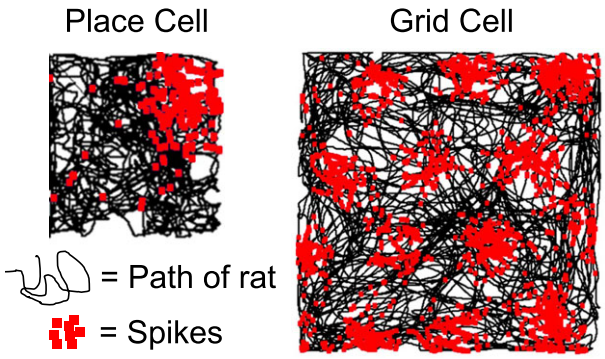

Figure 5. A place cell's firing field (left) and a grid cell's grid of firing fields (right) as a rat moves around an enclosure. Recorded by Elizabeth Marozzi. From [172]. © authors. (Online version in colour.)

Third, how different animals navigate the 3D world may be very different as well. Rodney Brooks explains how insects inspired his rejection of 3D models in robotics: 'Look at an insect, it can fly around and navigate with just a hundred thousand neurons. It can't be doing this very complex symbolic mathematical computations. There must be something different going on.' [159]. See also work by Barbara Webb [160–163] as well as [164]. Recent work at the intersection of insect navigation, computation, and robotics includes work on bees [165,166], flies (Fly-Net: [167]), and ants [168–170].

(a) Cognitive maps

However, the key paradigm for animal navigation over the past 120 years has been rodent navigation in mazes. The hope is that it will teach us about mammal navigation in general, and potentially provide important insights about how humans navigate the world. Navigation had its 'cognitive revolution' a decade before the rest of psychology when Tolman [171] rejected behaviourism, arguing that the relationship between the stimulus (maze) and the rat's response was mediated by the rat constructing a 'cognitive map' of the maze. This insight appeared to be confirmed by two findings that ultimately won the Nobel Prize in 2014.

First, in the early 1970s 'place cells' were found in the hippocampus. These cells fire when an animal is in a specific place in the environment (figure 5) [173]. In *The hippocampus as a cognitive map* [174,175], O'Keefe & Nadel argued that the hippocampus functioned as Tolman's cognitive map. Going beyond Tolman, they argue that this cognitive map had four properties. First, it is 'Euclidean' ('the metric of the cognitive map is Euclidean' [174]). Second, it is 'absolute' (it organizes or structures our experience: 'the brain must come equipped to impose a 3D Euclidean framework on experience' [175]). Third, it is 'world-centred' or 'allocentric' ('a non-centred stationary framework through which the organism and its egocentric spaces move' [174]. And fourth, it is 'innate' ('this framework is part of the innate machinery of the organism' [174]). Whether this 'Euclidean' ideal is borne out by the subsequent data is the key concern of this section.

Second, in the mid-2000s 'grid cells' were found in the entorhinal cortex. These cells fire when an animal crosses one of the cell's 'firing fields': locations in space arranged in a hexagonal grid that cover the entire area (figure 5)

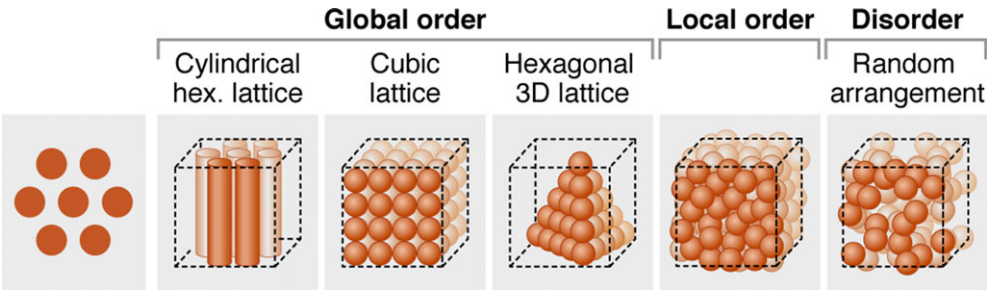

Figure 6. Hypotheses being tested by [11] and [12]. On the left is the 2D hexagonal grid of grid cell firing fields we saw in figure 5. The remaining panels explore potential 3D grid arrangements. Some sort of 3D 'global order' was originally hypothesised [193–197]. But Ginosar *et al.* [11] only find evidence of 'local order' in bats, whilst the results in rats [12] are consistent with a 'random arrangement'. From Ginosar *et al.* 2021 [11]. © authors. (Online version in colour.)

[176,177]. Each grid cell's hexagonal grid has its own scale (spacing between the firing fields), orientation (rotation of the firing fields), and phase (grid shifted in a certain direction). When grid cells were found, it was immediately hypothesized that their regular geometry could provide a metric map of the environment [178]: a map that preserves both the angles and distances between points, enabling path integration (for distance estimation) and vector navigation (for shortcuts) [179]. As Grieves & Jeffery [180] note: 'Given their mathematical and geometric properties it is almost hard to believe these cells exist at all'.

(b) Distortions in place cells and grid cells

However, the spatial mappings of both place cells and grid cells are subject to significant distortions [181,182], challenging the notion that they provide a metric map of the environment.

First, grid cells are loosely anchored to the environment, with a grid cell's firing fields being stable across time. But what this means is that if the environment is artificially expanded or contracted, the grid cell's firing fields will also be expanded or contracted by roughly 50%, creating a distorted geometry and metric to the space [183,184]. Similar distortions are also seen in place cells [185].

Second, grid cell firing fields are distorted by the borders of the environment [186], and there's evidence that these distortions impact spatial memory [187]. For instance, in a square environment these distortions lead to an elliptical grid [188], whilst in an irregular environment, like a trapezoid, the grid is almost obliterated [189,190]. As Krupic *et al.* [189] conclude:

> These results challenge the idea that the grid cells system can act as a universal spatial metric for the cognitive map as grid patterns change markedly between enclosures and even within the same enclosure.

Again, similar distortions are also seen in place cells [190].

Third, Boccara *et al.* [191] found that rewards also distort the regular grid arrangement of grid cell firing fields:

> Many grid fields moved toward goal locations, leading to long-lasting deformations of the entorhinal map.

Put simply, the entorhinal cortex (grid cell) cognitive map is attracted to goals: 'This demonstrates the influence of non-geometrical cognitive factors on the grid structure itself.' By contrast, Butler *et al.* [192] didn't find the grid distortion towards rewards that Boccara *et al.* [191] reported, although they did find that rewards affected the arrangement of grid cell firing fields in other ways (grids were translated and rotated, more closely spaced, less elliptical), and also grid cells fired more closer to rewards.

(c) Grid cells in 3D

Entering our meeting, the great unanswered question was how grid cell firing fields are organized in 3D space? They are regularly ordered on a 2D surface (e.g. the floor of an enclosure) (figure 5), and a number of models suggested that this regular 2D grid was likely a cross-section of a regular 3D grid in 3D space, where each firing field had a fixed angle and distance from one another ('global order' in figure 6) [193–197]. As Finkelstein *et al.* [197] noted in 2016: 'An important future test for this idea would be to record from 3-D grid cells: Do they support the notion of a metric representation of 3-D space?'

It was our privilege to have Gily Ginosar and Kate Jeffery present the two landmark studies on this very question at our meeting; Ginosar on bats flying in 3D space [11], and Jeffery on rats navigating a 3D maze [12]. The surprising answer from both studies is that grid cell firing fields are not arranged in a regular 3D grid, and so have no *global* 3D order [198]. In bats the grid cell firing fields 'exhibited only local order, creating a locally ordered metric for space' [11], with firing fields having similar spacing but not similar angles, whilst in rats firing fields were consistent with a random arrangement (no similarity in spacing or angles) [12].

The common finding that grid cell firing fields do not have a global 3D structure is especially striking given that bats and rats diverged evolutionarily 65 million years ago. And it challenges the assumption that grid cells provide a general purpose metric for space:

> …suggestions that grid cells are involved in geometric computations … were motivated by the highly geometric, periodic representation of 2D space by grid cells. Given our findings on the absence of global periodicity in 3D, it seems less plausible that 3D grid cells are involved in general purpose geometric computations… (Ginosar *et al.* [11])

Indeed, Grieves *et al.* [12] go further, and suggest that the structure of grid cell firing fields may reflect 'affordances' (potential for action). Explaining the difference between their study on rats and Ginosar *et al.* [11]'s study on bats, Grieves *et al.* [12] suggest that it is unlikely to arise from differences between the two species, but instead is likely to be 'due to how movement patterns through the volumetric spaces can affect grid self-organization.' For instance, bats could fly in any direction, whilst rats were constrained by the structure of the 3D maze they were navigating. As Kate Jeffery [199] explains, on this account 'the cognitive map is not fixed and rigid, like an artificial map', but instead provides 'a more flexible spatial code in which the map is

adapted to the movement possibilities of the space.' Kate Jeffery [200] continues the discussion in this issue, considering asymmetries and distortions in the firing fields of place cells, grid cells and head-direction cells.

(d) Cognitive maps and NeuroAI

In another contribution to this volume, Ida Momennejad [201] considers the convergence of neuroscience and machine learning (NeuroAI) (see also [202–205]). Nowhere has this convergence been more apparent than in navigation research. But how should we connect the neural representations (place cells and grid cells) that we are discussing with the computational models of scene understanding (SLAM and reinforcement learning) that we discussed in the computer vision section?

1. SLAM and place cells: One of the earliest models of SLAM (simultaneous localization and mapping) in computer vision was Maja Matarić's 'Navigating with a rat brain: a neurobiologically inspired model for robot spatial representation' [206]. Inspired by hippocampal place cells, Matarić aimed for a 'topological' rather than Euclidean map of the environment. Like the London tube map, 'topological' maps preserve the relationship between landmarks, but not their metric distances or directions. In Matarić's case, her robot could follow the perimeter of its enclosure, but its scene understanding was limited to which landmarks it would encounter in which order. By contrast, later models aimed to provide metric scene representations, with metric information either the explicit input [207] or the implicit output [208, p.107], [209]. Halfway between them, RatSLAM [210–213] aimed for a fine grained topological map that could facilitate shortcuts, but was less than fully metric ('the map does not follow a strict Cartesian coherence' [210]).

2. Reinforcement learning and grid cells: By the time grid cells were discovered in 2004–5, the literature's focus had largely shifted to reinforcement learning (learning navigation strategies, rather than hard-coding them; cf. [214]). Reconciling reinforcement learning with place and grid cells has been a key focus of the recent literature (for review see [215]), with both Euclidean and non-Euclidean approaches.

Euclidean: Banino *et al.* [216] and [217] show that grid cell-like patterns can spontaneously emerge when neural networks are trained to perform path integration. Banino *et al.* [216] incorporated this 'grid network' into a deep reinforcement learning agent, and further showed that this agent's performance had (unlike a comparison agent) all the hallmarks of a Euclidean spatial metric, such as distance estimation and 'vector navigation' (shortcuts to remembered locations [179,218,219]). They therefore conclude that grid cells, in both biological systems and machines, 'furnish agents with a Euclidean spatial metric'.

> …we argue that grid-like representations furnish agents with a Euclidean geometric framework – paralleling the proposed computational role in mammals as an early developing Kantian-like spatial scaffold that serves to organize perceptual experience… (From Banino *et al.* [216].)

Non-Euclidean: By contrast, at our meeting and in [220], Ida Momennejad argues that it's the non-Euclidean aspects of place and grid cells that need accounting for, such as the distortions of grid cell firing fields by rewards, as well as the fact that the majority of distance estimates in the hippocampus reflect path distance (taking into account obstacles) rather than direct Euclidean distance [221,222].

Reinforcement learning has been proposed as a way of capturing these non-Euclidean properties of place and grid cells. Gustafson & Daw [223] argue that an emphasis on path rather than Euclidean distance reflects the fact that place and grid cells 'are well adapted to support reinforcement learning', since efficient reinforcement learning requires that the inputs (place and grid cells) are already articulated in terms of the goals of navigation:

> Importantly, this exercise views the brain's spatial codes less as a representation for location per se, and instead [as] a *value function* over state space – a mapping of location to value.

More recently, Stachenfeld *et al.* [224,225] argue that place cells are the encoding of 'successor representations'. In reinforcement learning, 'successor representations' [226] provide a model of next steps, affording the flexibility of considering alternatives at each stage (versus model-free reinforcement learning), while avoiding the computational intractability of modelling the whole world (model-based reinforcement learning). On this account, place cells encode the likelihood that a location will be visited given the animal's current navigation strategy, explaining why locations that have the same path distance to a reward can still have different responses from the reward's place cell.

> Place cells in the hippocampus have traditionally been viewed as encoding an animal's current location. In contrast, the predictive map theory views these cells as encoding the animal's future locations. (Stachenfeld *et al.* [225])

Momennejad *et al.* [227], Russek *et al.* [228] and Geerts *et al.* [229] also suggest that successor representations capture the semi-flexible navigation strategies of humans and rodents.

Successor representations also promise to invert the traditional relationship between grid cells and place cells. Rather than grid cells being the neutral Euclidean input into place cells [230], Stachenfeld *et al.* [224,225] argue that grid cells are simply a higher-order abstraction (principal component analysis) of the successor representations captured in the place cell cognitive map (see also [231,232]).

(e) Vision and navigation

At our meeting we intentionally used '3D vision' in a broad (computer vision) sense to include 'vision for interaction with the 3D world' (e.g. navigation). But how should we relate the cognitive mapping of space in the hippocampus (place cells) and entorhinal cortex (grid cells) to the perceptual mapping of space in the visual cortex? Can navigation affect vision?

At our meeting Aman Saleem discussed his and colleagues' findings in [233] that the location of rewards alters firing rates in the mouse primary visual cortex as well as the hippocampus, suggesting an influence of cognitive maps all the way down to the earliest retinal maps in the cortex. Saleem *et al.* [233] therefore conclude that:

> …visual responses in V1 [primary visual cortex] are controlled by navigational signals, which are coherent with those encoded in hippocampus … The presence of such navigational signals as early as a primary sensory area suggests that they permeate sensory processing in the cortex.

Table 2. Models of human 3D vision.

Models of Human 3D Vision
No 3D Model
1. Direct Perception
2. Sensorimotor
3D Models that Recover Metric Scene Properties
3. Constraints (Non-Probabilistic)
4. Non-Linear Cue Combination (Strong Fusion)
5. Linear Cue Combination (Weak Fusion)
6. Inverse Graphics (Generative Models)
7. Deep Learning (Discriminative Models)
3D Models that Don't Recover Metric Scene Properties
8. Minimal Model
9. Fragmentation
10. Topology
11. Affine Geometry
12. Intrinsic Constraint
13. Perspective Space
14. Tripartite Model
15. Affordances / Embodiment
16. Task Dependence

The effect of self-motion on vision in mice and humans is further explored by Aman Saleem and colleagues (Horrocks, Mareschal & Saleem [234]) in this issue.

3. Human vision

Like computer vision and animal navigation, competing interpretations of human 3D vision alternate between adopting no 3D model, a 3D model that recovers metric scene properties, or a 3D model that doesn't recover metric scene properties. These different approaches are summarized in table 2.

(a) No 3D model

1. Direct Perception: The most famous 'no model' account of human vision is James Gibson's 'direct perception' [144,235,236], according to which we directly perceive the structure of the world through 'invariants' in the constantly changing retinal image:

> …*perceiving* is a registering of certain definite dimensions of invariance in the stimulus flux … The invariants are invariants of structure [236, p.249].
>
> The theory of the extracting of invariants by a visual system takes the place of theories of 'constancy' in perception, that is, explanations of how an observer might perceive the true colour, size, shape, motion, and direction-from-here of objects despite the wildly fluctuating sensory impressions on which the perceptions are based. [236, p.311].

But Gibson's proposal has been given a variety of interpretations by the contemporary literature. Wagner and colleagues [237] suggest that 'Gibson's … doctrine of realism implies that visual space should be strictly Euclidean … ' Warren [238] advances an 'affine' interpretation: 'humans do not in fact recover Euclidean structure—rather, they reliably perceive qualitative shape (hills, dales, courses and ridges), which is specified by the second-order differential structure of images.' Finally, Tsao & Tsao [239] argue for a 'topological' approach.

2. Sensorimotor: Sensorimotor (pixel-to-action) accounts of human vision have also been influential [240]. We saw sensorimotor (pixel-to-action) accounts previously in computer vision with Sergey Levine's work on reinforcement learning. In this issue Andrew Glennerster [241] argues that model-free reinforcement learning (without a 3D model) is a good model for human 3D vision, and his paper aims to show 'how a policy network could support the same behaviour as a system that uses a 3D reconstruction of the scene.'

(b) 3D models that recover metric scene properties

3. Constraints (Non-Probabilistic): The dominant approach to human and computer vision in the 1970s–80s was specifying the physical constraints on how the retinal image was produced, so that the inverse optics question (what 3D scene produced this 2D image?) had a unique solution: 'the resulting operation is defined uniquely by constraints it has to satisfy' (Marr [17, p.23]). However, in the late 1980s and early 1990s, it became apparent that simply specifying constraints would not suffice, although some argue that we simply have too limited a notion of these constraints [242,243].

4–6 (Below). Bayesian approaches: Instead, the visual system would have to decide which of the remaining potential percepts were more or less likely, framing perception as a probabilistic process: 'The principle aspect of this approach is the probabilistic representation of constraints.' (Clark & Yuille [244, p.218]). Bayesian models have been the dominant approach to human 3D vision for the past 25 years [245–248], and can be articulated in one of three ways:

4. Linear cue combination (weak fusion): The leading approach to human 3D vision [246,249] treats 3D vision simply as a problem of eradicating of sensory noise. It breaks 3D vision down into a series of 'cues' (e.g. stereo vision, motion parallax, structure from motion, perspective, shading), and assumes each cue gives an accurate (undistorted, unbiased) but imprecise (vague, noisy) depth estimate. It then reduces the effect of sensory noise by taking a weighted average of the individual cues: the less noisy a cue is, the more weight its estimate is given. Empirical support can be found in [250–261].

However, there are two key concerns with this approach. First, it assumes each cue gives an unbiased (undistorted) depth estimate. But we will see below that this isn't the case. And Domini & Caudek [262] argue that if 'the estimates of the world properties are biased, … it is meaningless to maximize reliability' (cf. [260]). Second, a significant number of studies are inconsistent with linear cue combination's prediction that the less noisy a cue is, the more weight its estimate is given [263–268] (see also [269] and [270]'s related methodological concerns).

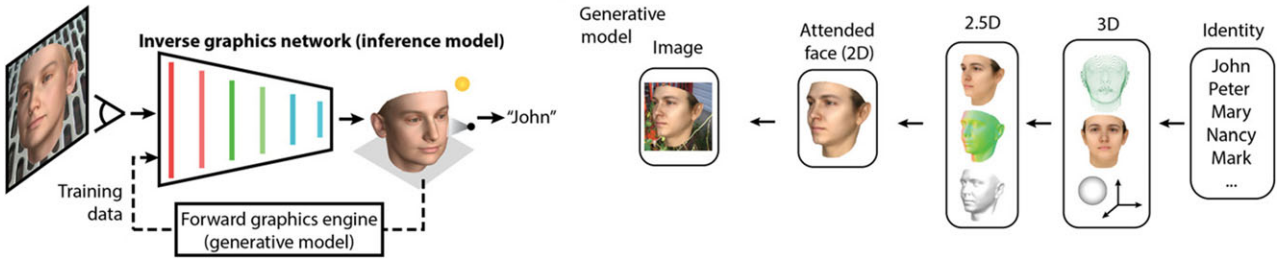

Figure 7. Inverse graphics network from [281]. How do we recognize someone as 'John'? Rather than train a neural network to directly identify people (2D image → identity), [281] use a 'generative model' (on the right) (identity → 3D scene → 2D image) to produce training data for an 'inference model' (on the left) that first estimates the 3D scene properties of the 2D image (2D image → 3D scene), before using this 3D scene estimate to identify the person (3D scene → identity). © authors. (Online version in colour.)

5. Nonlinear cue combination (strong fusion): By contrast, nonlinear cue combination assumes some degree of bias in the individual cues, and is really about 'the constraints needed to solve sensory information processing tasks, rather than just a method for reducing the effects of sensor noise.' (Clark & Yuille [244, p.222]). The focus is still on metric scene recovery, so that the cue combination rules 'inverting the world-image mapping are sufficient and, most importantly, *valid*.' (Clark & Yuille [244, p.222]). Nonlinear cue combination was originally more popular in the early 1990s, and ranges from the addition (rather than averaging) of cues [271,272], through to highly sophisticated interdependencies between cues [244,273].

6. Inverse graphics (generative models): Other, more recent, Bayesian accounts have done away with 'cues' altogether, and ask: 'what arrangement of lights, surfaces, and materials would give rise to this specific 2D image?' [274–280]. This 'inverse graphics' approach relies on 'analysis by synthesis': simulating 2D images of different 3D scenes to see under which 3D scene configuration the actual 2D image is most likely. Yildirim *et al.* [281] (figure 7) divide this into two stages, where a 'synthesis' model ('generative model', on the right in figure 7) simulates 2D images of 3D scenes, which are then used to train a separate 'analysis model' ('inverse graphics network', on the left in figure 7).

7. Deep learning (discriminative models): In contrast to 'inverse optics' models (3–6 in table 2), deep learning models can be thought of as reflecting a 'statistical appearance model' [282–284]: 'rather than learning the mappings between image quantities (cues) and physical quantities, we learn to represent the dimensions of variations within and among natural images, which in turn arise from the systematic effects that distal properties have on the image.' (Fleming & Storrs [284]). See also the discussion of the 'generative' versus 'discriminative' approaches in [285]. Strictly speaking, this approach aims to capture a compact representation of the variables that determine image structure (latent variables), rather than a representation of physical scene structure itself. However, as Fleming & Storrs [284] suggest, one may lead to the other: 'we may end up with internal representations that are well suited for describing the *distal* scene factors that have created those images.'

(c) Failures of metric scene reconstruction

Models 3–7 in table 2 provide different 'normative' or 'ideal observer' models [286] of how 3D vision 'ought' to act if it were trying to estimate the metric properties of the scene.

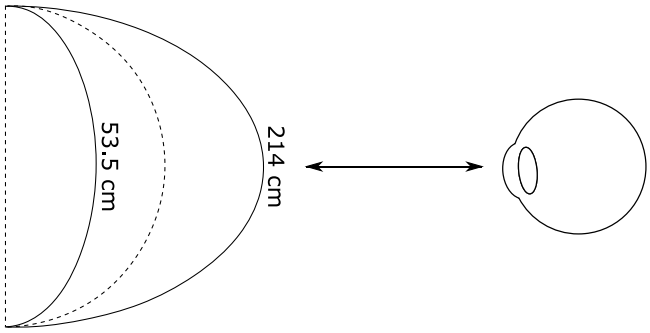

Figure 8. Schematic of the task in [287]. Participants set the depth of the cylinder so that its depth appeared to be proportional to its height (the dotted line). At near viewing distances (53.5 cm) the cylinder they produced was compressed in depth, while at far viewing distances (214 cm) the cylinder they produced was elongated in depth.

However, there's two ways in which human 3D vision regularly fails to live up to this standard of metric scene reconstruction.

1. Constancy (distortions): First, our depth perception is subject to considerable distortions. One vivid illustration is stereo vision (depth from disparity). In Johnston [287], participants viewed a cylinder whose depth was defined by stereo vision (disparity) alone. Their task was simple. Set the cylinder's depth so it was proportional to its height (the dotted line in figure 8). But the cylinder produced varied drastically in depth with the viewing distance. At 53.5 cm the cylinder they produced was compressed in depth, suggesting that depth from disparity is accentuated at close distances, while the cylinder they produced at 214 cm was elongated in depth, suggesting that depth from disparity is compressed at far distances. And these 3D shape distortions persist even in the presence of other cues [288,289].

Similarly, we seem to experience less depth when we look at the world with one eye closed, but there's little justification for why this should be the case if vision is estimating metric scene properties using the available information [290,291, pp.13–16].

Domini & Caudek [262] rightly consider 'the large failures of constancy of 3D metric structure over changes in viewing distance and/or orientation to be an important empirical finding that needs to be addressed by any theory'. First, metric cue integration accounts try to suggest that 'failures to observe depth constancy may be due to the influence of unmodelled flatness cues such as blur and accommodation' [249]. But this doesn't seem to fully meet the challenge. Second, others try to suggest that these distortions merely affect our visual

experience of depth (qualia), but not our metric scene estimates [291, pp.13–16]. But our visual experience of depth is exactly what a theory of 3D vision should explain. Third, some try to suggest that these distortions are a way of the visual system conveying how reliable our metric scene estimates are [292]. But this is not supported by the data [293].

2. Consistency (conflicts): Second, our judgements about visual space are often marked by pervasive inconsistencies. Di Luca *et al*. [294] find inconsistencies between estimates of depth, slant and curvature. Koenderink [295] finds inconsistencies between global and local depth judgements. Loomis *et al*. [296–298] find inconsistencies between absolute distance judgements and relative depth judgements. Koenderink *et al*. [299] find conflicting judgements of fronto-parallels. And Svarverud *et al*. [300] find visual space can be 'broken', with no consistent ordering between objects. Illusions also provide insights into inconsistency: Gillam & Chambers [301] find that position and size are inconsistent in the Müller-Lyer illusion, while Smeets *et al*. [302] find that the perceived centre of the Judd and Poggendorff illusions depends on the order in which the points are constructed. There's also a debate over whether we experience conflicting shape percepts when we see a rotated coin both as a circle and as an ellipse (Morales *et al*. [303–305] vs Linton [306] vs Burge & Burge [307]).

(d) 3D Models that don't recover metric scene properties

On 'normative' or 'ideal observer' models, these failures of metric scene reconstruction reflect a failure of evolution to live up to our rational standards of what vision 'ought' to be doing (Landy *et al*. [249]):

> Of course, there remains the possibility that we have characterized the sensory information *and the task* correctly, but the nervous system simply has not developed the mechanisms for performing optimally (Landy *et al*. [249], emphasis added)

By contrast, for models 8–16 in table 2, these failures of metric scene reconstruction suggest that we haven't specified the task correctly. Instead, these failures give us an important insight into the very different task that the visual system has set itself.

So a number of authors, including three of the present authors (Linton, Vishwanath, Domini), see the failures of metric scene reconstruction as a reason to question whether human 3D vision is trying to extract the metric 3D properties of the environment in the first place. Vishwanath [308] argues that 3D vision is '*the presentation* of causally efficacious visual information rather than an *inference* to objective external reality.' Domini & Caudek [262] argue that 'the goal of the visual system is to guarantee a successful interaction between the observer and the environment without recovering metric 3D (3D) information'. And Linton [309], p. 74 argues that 3D vision 'operates purely at the level of *phenomenal geometry*, and makes no claims about the *physical geometry of the physical world*'.

But if human 3D vision isn't trying to estimate the metric properties of the environment, what else could it be doing? The following models provide nine distinct alternatives.

8. Minimal model: In this issue, Paul Linton [310] argues for a 'minimal model' of 3D vision.

First, in order to answer the challenge of 'inconstancy', Linton decouples stereo vision from estimating scene properties such as distance and shape, rejecting 'triangulation'-based accounts of stereo vision that date back to Kepler [311] and Descartes [312]. Instead, on his account, stereo depth is simply a solution to a different (and entirely internal) problem, the eradication of rivalry between the two retinal images. Indeed, rather than the resulting 'inconstancy' being a problem to be solved, Linton argues that it's primarily through this 'inconstancy' that we judge size and distance.

Second, in order to answer the challenge of 'inconsistency', Linton argues that stereo vision and non-stereo cues (and therefore the inconsistencies between stereo vision and non-stereo cues) operate at different levels, with stereo vision affecting our perception (visual experience) of depth, whilst non-stereo cues (such as motion, perspective and shading) merely affect our cognition (judgements) of depth.

9–13 (Below). Qualitative models of scene geometry: The following five approaches suggest that human 3D vision captures scene geometry in a qualitative (fragmentary and/or distorted) sense, but differ as to the exact way in which scene geometry is fragmented and/or distorted. Often this takes the form of asking how loose (or permissive) the mathematical transformation from physical space to visual space is? From more to less permissive: Topology → Projective Geometry (Perspective) → Affine Geometry → Euclidean Geometry.

9. Fragmentation: A common response to conflicting depth estimates of the same 3D scene is to say that humans are only capable of local, and often inconsistent, depth judgements. For instance, Koenderink [295] suggests:

> Observers are quite content to live with any number of mutually inconsistent fragmentary representations since they can blindly depend on the consistency of the physical world.

And illusions lead [301,313] and [302] to reach a similar conclusion. Asking 'Does visual space exist?', Smeets *et al*. [302] conclude:

> It might be more fruitful to abandon the concept of a geometrically consistent perceptual (or motor) space altogether. Instead, one can regard perception as a set of independent local estimates of various spatial attributes.

However, Koenderink *et al*. [314] raise the possibility that while depth judgements are fragmentary, they're still consistent, arguing that you can build up a coherent global map from observers' local depth judgements. But Koenderink *et al*. [314] argue that observers are unable to build up a coherent global map for themselves since they experience visual space neither globally nor locally, but as a patchwork of 'hills' and 'troughs':

> Possibly the data structure itself is not a whole, but rather a quilt of locally coherent, but mutually only weakly synchronized patches.

10. Topology: Nowhere has the shift in thinking about human 3D vision from a quantitative 'inverse optics' approach to a qualitative approach been more apparent than in 'shape-from-shading'. Initial research in the 1980s tied human vision closely to computer vision [315]. Now leading figures from that tradition seek to sharply distinguish human vision from 'inverse optics'. So, Jan Koenderink, Andrea van Doorn, and colleagues suggest [316]:

> It may well be the case that the whole notion of shape from shading is spurious … , and that biological vision research should leave it to computer vision engineers.

Similarly, Steven Zucker, in 'On qualitative shape inferences: a journey from geometry to topology' (2020) [317] argues:

> …we argue that the perception of shape is qualitative, not quantitative, a point that has been well understood in visual psychophysics for decades. This suggests that we should not be

seeking to solve the shape-from-shading equations, but should look for qualitative (that is, topological) solutions instead.

And both Koenderink *et al*. and Zucker point to the considerable differences not just between observers' 3D estimates and reality, but also between observers themselves.

For Kunsberg & Zucker [318,319], this qualitative approach is best expressed through 'topology', the study of what remains true about a surface after it has been distorted in an arbitrary way. For instance, if we think about deforming a mesh, neighbouring points on the mesh before deformation will still be neighbouring points afterwards. For Zucker, the topology of the surface that emerges in shape from shading is anchored in a few places by 'critical contours' (such as peaks and troughs), which form 'a kind of scaffold on which the shape can be readily built.' But beyond that, each person's interpretation of the surface is largely subjective and unconstrained.

11. Affine geometry: By contrast, Koenderink *et al*. [320] suggest observers can recover something more than topology, but still less than Euclidean geometry, noting that observers are 'surprisingly close to the physical layout modulo a gauge transformation, with the deviations being mainly an isotropic rotation and scaling.' This is 'affine geometry', which captures the Euclidean geometry subject to a homogeneous stretch or shear. Affine geometry is more structured than topology, preserving the relationship between parallel lines, but not the distance or angles between points.

As we saw earlier in computer vision, the information from both stereo vision (disparity) and structure from motion also lends itself to analysis at the affine level. But evidence of whether humans can recover affine structure from stereo vision or motion is equivocal. Domini *et al.* [321] find that structure from motion is not affine, while [322,323] find that structure from motion is affine but stereo vision is not. Todd *et al.* [324] did however find that whilst the mapping from physical space to visual space is not affine for stereo vision, stereo vision is at least internally consistent in an affine sense ('perceptions had an internally consistent affine structure'), since line bisections in different directions were consistent with one another.

More broadly, Wagner *et al.* [325] suggest that size and distance judgements are best thought of as affine transformations of physical space, and Glennerster *et al*. [326] that pointing errors in virtual spaces reflect affine distortions. Affine geometry also played a key role in early versions of Domini & Caudek's 'intrinsic constraint' cue combination model [262,327]: 'Our main claim … is that the brain extracts from retinal signals the local affine information of environmental objects.'

12. Intrinsic constraint: In this issue Fulvio Domini [328] presents a new version of the 'intrinsic constraint' model. On this new account, perceived depth is still linearly related to (and therefore still an affine transformation of) physical depth. But what this new 'intrinsic constraint' model is trying to do is not estimate affine depth *per se*, but simply maximize the 3D signal in the image (while minimizing nuisance variables, such as viewing conditions and materials). To achieve this, the 'intrinsic constraint' model uses a 'vector sum' that adds (rather than averages) the depth estimates from the individual cues, so that the more cues you have (and the more 3D signal is in the image), the more depth you see. While this model aims to achieve a more stable representation across viewing conditions, it also explains why adding or removing cues can lead to inconsistent depth estimates.

13. Perspective space: Train tracks appear to converge as they recede in distance, suggesting that instead of an affine transform (which preserves parallel lines), visual space is a perspective projection of Euclidean space (with visual space converging to a vanishing point) [329–333]. Wagner *et al.* [237] find that human judgements correspond better to a perspective projection than an affine transform. But what's so surprising about the perspective space account is just how shallow visual space appears to be. Erkelens [332] asked participants to match the perceived convergence of railway lines using compasses, and the vanishing point inferred was no further than 6 m, suggesting that visual space is compressed in depth, like a bas relief, to fit within 0–6 m of physical space.

14. Tripartite model: Perhaps we should think of visual spaces, rather than one single visual space. In this issue Dhanraj Vishwanath [334] argues for a 'tripartite model' of visual space, according to which there are three 'distinct and dissociated encodings' for (a) 3D shape, (b) 'egocentric' (observer to object) distances, and (c) 'exocentric' (object to object) distances. However, often viewing conditions will only support one or two of these encodings, explaining the 'inconsistencies' reported in the literature above.

Vishwanath argues for these three distinct encodings based on their different experiential 'qualities', in contrast to standard models that 'typically do not make a fundamental distinction among these different modes of spatial experience'. For instance, the dual nature of pictures reflects the fact that pictorial space supports 3D shape perception but not 'egocentric' and 'exocentric' distance perception, whilst the vivid separation in depth typically associated with stereo vision (depth from disparity) reflects the 'exocentric' distance encoding.

15. Affordances/Embodiment: The past couple of decades have seen the rise of a 'pragmatic turn' in cognitive science, according to which 'cognition should not be understood as providing models of the world, but as subserving action' [335–337]. This has led to increased interest in the visual processing of 'affordances' (potential for action) [236,338]. We have already encountered one such affordance (traversability) in our discussions of robot navigation [104] and rats navigating 3D mazes [12]. And Sarah Creem-Regehr's [339] talk at our meeting focused on affordances in virtual and augmented reality.

But what are the implications for 3D vision? Do affordances replace our perception of 3D surfaces, as Sergey Levine [340] suggests they ought to for robotics? In some passages Gibson suggests so ('What animals need to perceive is not layout as such but the affordances of the layout', Gibson [341, pp.157–158]), but contemporary Gibsonians reject this (Warren [342]).

Instead, the closest we come to this view are 'embodied' theories of perception [343,344] that claim that 'explicit awareness of spatial layout varies not only with relevant optical and ocular-motor variables but also as a function of the costs associated with performing intended actions.' The classic claim is that hills are perceived as steeper when you wear a heavy backpack or if you are elderly ([345]; see also [346–348]). However, this theory has been criticized from both experimental [349] and theoretical [350,351] perspectives, with the suggestion being that the effect of the rucksack is a consequence of participants' trying to 'act as they should' in the experiment. For the latest iteration of this debate see [352–355].

16. Task dependence: Still, perhaps there is something to the idea of tying scene estimates to the task being performed. Indeed, the suggestion that 3D vision is task-dependent is

one of the most common responses to the 'inconsistencies' in 3D vision outlined above. For example, Glennerster *et al.* [356] found that depth constancy for stereo vision (disparity) depends on the nature of the task, and the same was found for tasks involving both stereo vision and motion parallax [357]. Norman *et al.* [358] tested the relationship between objects in action space and found that: 'Whether a Euclidean or affine compressed visual space was obtained depended not upon any characteristic of the visual stimuli, but upon the specific task employed by the observer.' Wagner & Gambino [325] 'embrace the idea that visual space is a living, malleable entity whose geometry changes with experimental conditions and shifts in observer attitude.' Finally, Warren [338] concludes that 'there *is* no consistent visual space. Rather, perception by an active agent is task-specific and information-driven, such that judgements of different properties of layout are based on different optical variables.'

But task dependence doesn't necessarily mean giving up on the concept of visual space. Mel Goodale and David Milner have long argued that task-dependence implies two models of visual space, one for conscious visual perception and the other for action [359–361]. A key claim of this account is that vision for perception is distorted by illusions, whilst vision for action is not. But an alternative explanation for this effect is that eye movements are different in perception and action, explaining the different effect of illusions. So, in this issue, to rule this out, Whitwell *et al.* [362] show that the effect still persists even when there are no significant differences between eye movements in perception and action.

(e) Virtual reality

Virtual reality enables us to study perception in a more ecologically valid way that is closer to the real world than conventional displays. But it also enables us to test vision in a more ecologically invalid way by decoupling the visual and physical consequences of our actions. This technique is used in two papers in this issue. First, Horrocks, Mareschal, & Saleem review [363] how virtual reality is being used to study the effects of locomotion on optic flow in humans and mice. Second, Maselli, Ofek, Cohn, Hinckley & Gonzalez-Franco [364] test how participants respond to displacing the location of a virtual (seen) hand relative to their physical (unseen) hand as they reach for an object, and find more efficient corrections towards the body midline.

However, we are still in the process of understanding the limitations of virtual reality. At our meeting, Douglas Lanman outlined the progress that his Display Systems Research team at Meta (Facebook) Reality Labs is making towards the 'visual Turing test': creating a display indistinguishable from reality [365,366]. But he also cautioned against equating virtual reality with real-world vision given the optical distortions that exist in virtual reality displays [367–369]. The concern that virtual reality doesn't reflect real-world perception is also shared by two papers in this issue. First, Creem-Regehr, Stefanucci & Bodenheimer [339] show how distances are underestimated in virtual reality, and the strategies that can be used to improve distance perception. Second, Rzepka, Hussey, Maltz, Babin, Wilcox & Culham [370] find that participants rely far more on the familiar size of objects when making distance judgements in virtual reality than they do in the real world.

(f) 3D space and visual impairment

The final two contributions to this issue explore how visual impairments affect our perception of space, and the ability of the human brain to adapt to these impairments.

First, 8% of the general population appear to have no stereo vision (they are unable to extract depth from disparity) [371,372]. Poor stereo vision is known to affect fine motor skills (such as threading a bead on wire) [373], reaching and grasping [374], and walking across uneven terrain [375]. Sue Barry [376] ('Stereo Sue' in Oliver Sacks' *The Mind's Eye* [377]) and Bruce Bridgeman [378] provide vivid personal descriptions of how recovering stereo vision transformed their visual experience:

> Gaining stereovision, I thought, would augment my perception of depth but not change it in any fundamental way. So, I was completely unprepared for my new appreciation of space… (Barry [376], p.111)

Extrapolated to the world population, stereo vision deficits affect over half a billion people, so a key concern is better understanding what causes stereo vision deficits, how they affect our interactions with the world, and how they might be treated. In this issue Niechwiej-Szwedo, Colpa & Wong [379] review the effect that amblyopia (lazy eye) has on the development of reaching and grasping, documenting how young children lag behind their peers, and older children develop compensatory strategies.

Second, recovery from early blindness can cause very selective visual deficits. As Ione Fine and colleagues' found with Mike May when he recovered his sight, his perception of visual motion was relatively 'normal', whilst his perception of 3D form remained permanently impaired [380,381]. In this issue, Fine & Park explain this by pointing to the fact that early blind individuals use auditory motion for many of the tasks we would typically attribute to 3D vision (e.g. navigating a busy interaction using noise from passing cars). But how is this possible? Fine & Park [382] find that this ability relies on the auditory system adopting brain area hMT+, that's associated with visual motion processing in normal observers, but it changes the nature of the motion processing that hMT+ engages in to accommodate the low spatial resolution of auditory information.

4. Conclusion

As we noted at the beginning, the purpose of our meeting was to capitalize on a brief moment when computer vision, animal navigation, and human vision are all pausing and asking what the most appropriate representation for 3D vision and action really is? The argument of this article, and indeed of this issue, is that our understanding of how brains and computers do 3D vision is at a crossroads. As Andrew Glennerster [383], one of the contributors to this issue, notes, 'we're moving away from the idea that what the brain does is something complicated, which is easy for us to understand', namely a metric 3D map of the environment, 'toward the view that the brain does something which is easy for it to do, but really quite hard for us to understand'. This issue presents sixteen perspectives on what that 'something' might be. But it's meant to be the beginning of a conversation, not the end. And, at a time when neuroscience, and science in general, is thought to be

in the midst of a 'theory crisis'[2], our hope is to have put these theoretical questions back at the centre of 3D vision.

Data accessibility. This article has no additional data.

Authors' contributions. P.L.: writing—original draft, writing—review and editing; M.J.M. writing—review and editing; J.L.A.R.: writing—review and editing; D.V.: writing—review and editing; S.H.C.-R.: writing—review and editing; F.D.: writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. P.L. is supported by the Presidential Scholars in Society and Neuroscience, Columbia University and a Fellowship in Art, Humanities and Neuroscience at the Italian Academy for Advanced Studies in America, Columbia University. D.V. was supported by a Leverhulme Trust Research Project Grant (grant no. RPG-2016-269: 'What does it mean to 'See in 3D'?'). F.D. is supported by a National Science Foundation grant (NSF #BCS 2120610: 'A test of a novel non-probabilistic model of 3D cue integration').

Acknowledgements. We would like to thank (in alphabetical order) Andrew Davison, Ali Eslami, Andrew Glennerster, Gily Ginosar, Kate Jeffery, Nikolaus Kriegeskorte, Sergey Levine, Ida Momennejad and Christopher Peacocke for their comments on an earlier draft (errors and omissions remain our own). We would like to thank Velodyne Lidar (figure 1), SM Ali Eslami (figure 3), and Ben Mildenhall (figure 4) for permission to use their images, and Kate Jeffery (figure 5) and Gily Ginosar (figure 6) for providing hi-res images. Finally, we would like to thank Helen Eaton for acting as Senior Commissioning Editor and Alice Power for acting as Commissioning Editor, Simon Clackson for acting as Production Manager, Jessica Erland for proof-reading the issue, and Annabel Sturgess for acting as Senior Scientific Programmes Officer for the Scientific Meeting.

Endnotes

[1]Office of Naval Research grant ONR MURI N00014-16-1-2007. Multidisciplinary University Research Initiative (MURI): 'The goal of this MURI team is to develop machines that have the following capabilities: … Achieve deep understanding of scenes and events through joint parsing and cognitive reasoning about appearance, geometry, functions, physics, causality, intents and belief of agents'. Members (USA): Song-Chun Zhu, Martial Hebert, Abhinav Gupta, Nancy Kanwisher, Josh Tenenbaum, Derek Hoiem, Brian Scholl, Tao Gao, Noah D. Goodman. (UK): Philip Torr, Philippe Schyns, Andrew Glennerster, Ales Leonardis.

[2]See [384, 39:33–41:40], [385–387], [388, 3:05–4:45], [389–398].

References

1. Knight W. 2022 A New Trick Lets Artificial Intelligence See in 3D. *Wired*. See https://www.wired.com/story/new-way-ai-see-3d/
2. Jumper J *et al.* 2021 Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, Article 7873. (doi:10.1038/s41586-021-03819-2)
3. Tunyasuvunakool K *et al.* 2021 Highly accurate protein structure prediction for the human proteome. *Nature* **596**, Article 7873. (doi:10.1038/s41586-021-03828-1)
4. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2017 Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253. (doi:10.1017/S0140525X16001837)
5. Zhu Y *et al.* 2020 Dark, beyond deep: a paradigm shift to cognitive AI with Humanlike common sense. *Engineering* **6**, 310–345. (doi:10.1016/j.eng.2020.01.011)
6. LeCun Y. 2022 A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. *Open Review* **62**, 1–62.
7. Chen H, Venkatesh R, Friedman Y, Wu J, Tenenbaum JB, Yamins DLK, Bear DM. 2022 *Unsupervised Segmentation in Real-World Images via Spelke Object Inference* (arXiv:2205.08515). arXiv. (doi:10.48550/arXiv.2205.08515)
8. Andrew Davison [@AjdDavison]. 2021 *For me the hardest part of robotics is not learning action, but still how to make 3D scene understanding actually work robustly, precisely and efficiently with real sensors in the cluttered real world. #SpatialAI* See https://t.co/0MjvQfnoru [Tweet]. Twitter. https://twitter.com/AjdDavison/status/1476145469063671808
9. Andrew Davison [@AjdDavison]. 2021 *Not usually one to just agree with Elon, but I think he's saying the same thing here as I was in my last tweet: The hardest part of AI is perception: Going from real sensor data to an efficient, but explicit, scene representation—Then your robot can do pretty much anything.* [Tweet]. Twitter. See https://twitter.com/AjdDavison/status/1476234087979532291
10. Lex Fridman (Director). 2021 *Elon Musk: SpaceX, Mars, Tesla Autopilot, Self-Driving, Robotics, and AI | Lex Fridman Podcast #252*. See https://www.youtube.com/watch?v=DxREm3s1scA&t=4113s
11. Ginosar G, Aljadeff J, Burak Y, Sompolinsky H, Las L, Ulanovsky N. 2021 Locally ordered representation of 3D space in the entorhinal cortex. *Nature* **596**, Article 7872. (doi:10.1038/s41586-021-03783-x)
12. Grieves RM, Jedidi-Ayoub S, Mishchanchuk K, Liu A, Renaudineau S, Duvelle É, Jeffery KJ. 2021 Irregular distribution of grid cell firing fields in rats exploring a 3D volumetric space. *Nat. Neurosci.* **24**, Article 11. (doi:10.1038/s41593-021-00907-4)
13. Royal Society. 2021 *New approaches to 3D vision*. See https://royalsociety.org/science-events-and-lectures/2021/11/3d-vision/
14. Roberts L. 1963 Machine Perception of Three-Dimensional Solids. PhD thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering.
15. Roberts LG. 1965 Machine Perception of 3-D Solids. In *Optical and electro-optical information processing* (eds JT Tippett, D Berkowitz, LC Clapp, CJ Koester, A Vanderburgh), pp. 159–197. Cambridge, MA: MIT Press.
16. Huang TS. 1996 Computer vision: Evolution and promise. In *CERN European Organization for Nuclear Research-Reports-CERN*, pp. 21–26.
17. Marr D. 1982 *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.
18. Waltz D. 1975 Understanding line drawings of scenes with shadows. In *The psychology of computer vision* (eds PH Winston, B Horn), pp. 19–91. New York, NY: McGraw-Hill.
19. Horn BKP. 1985 The Role of Optics in Machine Vision. *Topical Meeting on Machine Vision (1985), Paper WA1*, WA1. See https://opg.optica.org/abstract.cfm?uri=MV-1985-WA1
20. Charniak E, McDermott D. 1985 *Introduction to artificial intelligence*. Boston, MA: Addison-Wesley.
21. Brooks RA. 1999 *Cambrian intelligence: The early history of the New AI*. Cambridge, MA: MIT Press.
22. Bajcsy R. 1985 *Presentation*. 2nd Workshop on Computer Vision, Bel Air, MI.
23. Bajcsy R. 1988 Active perception. *Proc. IEEE* **76**, 966–1005. (doi:10.1109/5.5968)
24. Ballard D. 1987 Eye Movements and Spatial Cognition. *University of Rochester, Computer Science, Technical Report, TR218*. See https://urresearch.rochester.edu/institutionalPublicationPublicView.action?institutionalItemId=5804
25. Ballard DH. 1991 Animate vision. *Artif. Intell.* **48**, 57–86. (doi:10.1016/0004-3702(91)90080-4)
26. Ballard DH, Brown CM. 1992 Principles of animate vision. *CVGIP: Image Underst.* **56**, 3–21. (doi:10.1016/1049-9660(92)90081-D)
27. Tenenbaum JM. 1970 Accommodation in computer vision. PhD Thesis, Stanford University, CA.
28. Brooks R. 1986 A robust layered control system for a mobile robot. *IEEE J. Robot. Autom.* **2**, 14–23. (doi:10.1109/JRA.1986.1087032)
29. Brooks RA. 1991 Intelligence without reason. In *Proc. of the 12th Int. Joint Conf. on Artificial Intelligence – vol. 1*, pp. 569–595. San Francisco, CA: Morgan Kaufmann Publishers Inc.
30. Blake A, Yuille AL. 1992 *Active vision*. Cambridge, MA: MIT Press.

31. Aloimonos Y. 1993 *Active perception*. Mahwah, NJ: Lawrence Erlbaum Associates.
32. Yarbus AL. 1967 Eye Movements During Perception of Complex Objects. In *Eye movements and vision* (eds AL Yarbus), pp. 171–211. Berlin, Germany: Springer US.
33. O'Regan JK, Lévy-Schoen A. 1983 Integrating visual information from successive fixations: does trans-saccadic fusion exist? *Vision Res.* **23**, 765–768. (doi:10.1016/0042-6989(83)90198-0)
34. Brooks RA. 1990 Elephants don't play chess. *Robot. Auton. Syst.* **6**, 3–15. (doi:10.1016/S0921-8890(05)80025-9)
35. Brooks RA. 1991*a* Intelligence without representation. *Artif. Intell.* **47**, 139–159. (doi:10.1016/0004-3702(91)90053-M)
36. Brooks RA. 1991*c* New approaches to robotics. *Science* **253**, 1227–1232. (doi:10.1126/science.253.5025.1227)
37. Faugeras O. 1995 Stratification of three-dimensional vision: projective, affine, and metric representations. *JOSA A* **12**, 465–484. (doi:10.1364/JOSAA.12.000465)
38. Faugeras OD. 1992 What can be seen in three dimensions with an uncalibrated stereo rig? In *Computer vision—ECCV'92* (eds G. Sandini), pp. 563–578. Berlin, Germany: Springer. (doi:10.1007/3-540-55426-2_61)
39. Beardsley PA, Zisserman A, Murray DW. 1997 Sequential updating of projective and affine structure from motion. *Int. J. Comput. Vision* **23**, 235–259. (doi:10.1023/A:1007923216416)
40. Koenderink JJ, Doorn AJv. 1991 Affine structure from motion. *JOSA A* **8**, 377–385. (doi:10.1364/JOSAA.8.000377)
41. Hartley R, Gupta R, Chang T. 1992 Stereo from uncalibrated cameras. In *Proc. 1992 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 761–764. Los Alamitos, CA: IEEE Computer Society Press.
42. Zeller C, Faugeras O. 1994 Applications of non-metric vision to some visual guided tasks. In *Proc. of 12th Int. Conf. on Pattern Recognition, 1*, vol. **1**, pp. 132–136. Los Alamitos, CA: IEEE Computer Society Press. See (doi:10.1109/ICPR.1994.576244)
43. Beardsley PA, Reid ID, Zisserman A, Murray DW. 1995 Active visual navigation using non-metric structure. *Proc. of IEEE Int. Conf. on Computer Vision*, pp. 58–64. Los Alamitos, CA: IEEE Computer Society Press. (doi:10.1109/ICCV.1995.466806)
44. Jacobs DW. 1994 Generalizing invariants for 3-D to 2-D matching. In *Applications of invariance in computer vision* (eds JL Mundy, A Zisserman, D Forsyth), pp. 415–434. Berlin, Germany: Springer.
45. Carlsson S. 1998 Geometric structure and view invariant recognition. *Phil. Trans. R. Soc. Lond. A: Math. Phys. Eng. Sci.* **356**, 1233–1250. (doi:10.1098/rsta.1998.0219)
46. Woodbury EJ, Congleton RS, Morse JH, Stitch ML. 1961 Design and operation of an experimental colidar. *IRE WESCON Conv. Record, 1961*.
47. Thrun S *et al.* 2007 Stanley: The Robot That Won the DARPA Grand Challenge. In *The 2005 DARPA grand challenge: the great robot race* (eds M. Buehler, K. Iagnemma, S. Singh), pp. 1–43. Berlin, Germany: Springer.
48. Abate T. 2007 *Robots, start your engines*. SFGATE. https://www.sfgate.com/bayarea/article/Robots-start-your-engines-3236937.php
49. Velodyne Lidar. 2017 It Began With a Race…16 Years of Velodyne LiDAR. *Velodyne Lidar*. See https://velodynelidar.com/blog/it-began-with-a-race16-years-of-velodyne-lidar/
50. Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems, 25*. See https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
51. Ohnsman A. 2013 Tesla CEO Talking With Google About 'Autopilot' Systems. *Bloomberg.Com.* See https://www.bloomberg.com/news/articles/2013-05-07/tesla-ceo-talking-with-google-about-autopilot-systems
52. Elon Musk [@elonmusk]. 2021 *@WholeMarsBlog Vision became so good that radar actually reduced SNR, so radar was turned off. Humans drive with eyes & biological neural nets, so makes sense that cameras & silicon neural nets are only way to achieve generalized solution to self-driving.* [Tweet]. Twitter. See https://twitter.com/elonmusk/status/1447588987317547014
53. Thrun S, Burgard W, Fox D. 2000 A real-time algorithm for mobile robot mapping with applications to multi-robot and 3D mapping. In *Proc. 2000 ICRA. Millennium Conference. IEEE Int. Conf. on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065), 1*, vol. 1, pp. 321–328. Los Alamitos, CA: IEEE Computer Society Press. (doi:10.1109/ROBOT.2000.844077)
54. Durrant-Whyte H, Rye D, Nebot E. 1996 Localization of Autonomous Guided Vehicles. In *Robotics research* (eds G Giralt, G Hirzinger), pp. 613–625. Berlin, Germany: Springer.
55. Durrant-Whyte H, Bailey T. 2006 Simultaneous localization and mapping: part I. *IEEE Robot. Autom. Mag.* **13**, 99–110. (doi:10.1109/MRA.2006.1638022)
56. Thrun S, Burgard W, Fox D. 2005 *Probabilistic robotics*. Cambridge, MA: MIT Press.
57. Newman P, Leonard J, Tardos JD, Neira J. 2002 Explore and return: Experimental validation of real-time concurrent mapping and localization. In *Proc. 2002 IEEE Int. Conf. on Robotics and Automation (Cat. No.02CH37292), 2*, vol. **2**, pp. 1802–1809. Los Alamitos, CA: IEEE Computer Society Press. (doi:10.1109/ROBOT.2002.1014803)
58. Smith R, Self M, Cheeseman P. 1990 Estimating Uncertain Spatial Relationships in Robotics. In *Autonomous robot vehicles* (eds IJ Cox, GT Wilfong), pp. 167–193. Berlin, Germany: Springer.
59. Montemerlo M, Thrun S, Koller D, Wegbreit B. 2002 FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Eighteenth National Conf. on Artificial Intelligence*, pp. 593–598.
60. Montemerlo M, Thrun S, Roller D, Wegbreit B. 2003 FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In *Proc. of the 18th Int. Joint Conf. on Artificial Intelligence*, pp. 1151–1156. San Francisco, CA: Morgan Kaufmann Publishers Inc.
61. Leonard JJ, Durrant-Whyte HF. 1991 Simultaneous map building and localization for an autonomous mobile robot. In *Proceedings IROS '91:IEEE/RSJ International Workshop on Intelligent Robots and Systems '91*, vol. **3**, pp. 1442–1447. (doi:10.1109/IROS.1991.174711)
62. Clark S, Dissanayake G. 1999 Simultaneous localisation and map building using millimetre wave radar to extract natural features. In *Proc. 1999 IEEE Int. Conf. on Robotics and Automation (Cat. No.99CH36288C), 2*, vol. **2**, pp. 1316–1321. Los Alamitos, CA: IEEE Computer Society Press. (doi:10.1109/ROBOT.1999.772543)
63. Guivant J, Nebot E, Baiker S. 2000 Autonomous navigation and map building using laser range sensors in outdoor applications. *J. Robot. Syst.* **17**, 3817–3822.
64. Davison AJ, Murray DW. 2002 Simultaneous localization and map-building using active vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 865–880. (doi:10.1109/TPAMI.2002.1017615)
65. Davison AJ, Murray DW. 1998 Mobile robot localisation using active vision. In *Computer vision—ECCV'98* (eds H Burkhardt, B Neumann), pp. 809–825. Berlin, Germany: Springer.
66. Davison AJ. 1998 Mobile Robot Navigation using Active Vision. PhD thesis, Oxford, UK: University of Oxford.
67. Davison AJ, Reid ID, Molton ND, Stasse O. 2007 MonoSLAM: real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 1052–1067. (doi:10.1109/TPAMI.2007.1049)
68. Davison. 2003 Real-time simultaneous localisation and mapping with a single camera. In *Proc. Ninth IEEE Int. Conf. on Computer Vision*, vol. 2, pp. 1403–1410. Los Alamitos, CA: IEEE Computer Society Press. (doi:10.1109/ICCV.2003.1238654)
69. Gupta S, Davidson J, Levine S, Sukthankar R, Malik J. 2017 Cognitive Mapping and Planning for Visual Navigation. In *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 7272–7281. (doi:10.1109/CVPR.2017.769)
70. Davison AJ. 2018 FutureMapping: The Computational Structure of Spatial AI Systems. *ArXiv:1803.11288 [Cs]*. See http://arxiv.org/abs/1803.11288
71. Davison AJ, Ortiz J. 2019 FutureMapping 2: Gaussian Belief Propagation for Spatial AI. *ArXiv:1910.14139 [Cs]*. See http://arxiv.org/abs/1910.14139
72. Sucar E, Liu S, Ortiz J, Davison AJ. 2021 iMAP: Implicit Mapping and Positioning in Real-Time. *ArXiv:2103.12352 [Cs]*. See http://arxiv.org/abs/2103.12352
73. Tung H-YF, Cheng R, Fragkiadaki K. 2019 Learning Spatial Common Sense With Geometry-Aware Recurrent Networks. In *2019 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pp. 2590–2598. (doi:10.1109/CVPR.2019.00270)

74. Harley AW, Lakshmikanth SK, Li F, Zhou X, Tung H-YF, Fragkiadaki K. 2020 Learning from Unlabelled Videos Using Contrastive Predictive Neural 3D Mapping. *ArXiv:1906.03764 [Cs]*. See http://arxiv.org/abs/1906.03764
75. Wald J, Dhamo H, Navab N, Tombari F. 2020 Learning 3D Semantic Scene Graphs from 3D Indoor Reconstructions. *ArXiv:2004.03967 [Cs]*. See http://arxiv.org/abs/2004.03967
76. Rosinol A, Abate M, Chang Y, Carlone L. 2020 Kimera: An Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. *ArXiv:1910.02490 [Cs]*. See http://arxiv.org/abs/1910.02490
77. Rosinol A, Violette A, Abate M, Hughes N, Chang Y, Shi J, Gupta A, Carlone L. 2021 Kimera: from SLAM to spatial perception with 3D dynamic scene graphs. *Int. J. Robot. Res.* **40**, 1510–1546. (doi:10.1177/02783649211056674)
78. Gupta S, Tolani V, Davidson J, Levine S, Sukthankar R, Malik J. 2020 Cognitive mapping and planning for visual navigation. *Int. J. Comput. Vision* **128**, 1311–1330. (doi:10.1007/s11263-019-01236-7)
79. Zhang J, Tai L, Liu M, Boedecker J, Burgard W. 2017 Neural SLAM: Learning to Explore with External Memory (arXiv:1706.09520). arXiv. (doi:10.48550/arXiv.1706.09520)
80. Parisotto E, Salakhutdinov R. 2017 Neural Map: Structured Memory for Deep Reinforcement Learning (arXiv:1702.08360). arXiv. (doi:10.48550/arXiv.1702.08360)
81. Gordon D, Kembhavi A, Rastegari M, Redmon J, Fox D, Farhadi A. 2018 IQA: Visual Question Answering in Interactive Environments (arXiv:1712.03316). arXiv. (doi:10.48550/arXiv.1712.03316)
82. Chen T, Gupta,S, Gupta A. 2019 Learning Exploration Policies for Navigation (arXiv:1903.01959). arXiv. (doi:10.48550/arXiv.1903.01959)
83. Chaplot DS, Gandhi D, Gupta S, Gupta A, Salakhutdinov R. 2020 Learning to Explore using Active Neural SLAM (arXiv:2004.05155). arXiv. (doi:10.48550/arXiv.2004.05155)
84. Henriques JF, Vedaldi A. 2018 MapNet: An Allocentric Spatial Memory for Mapping Environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8476–8484. (doi:10.1109/CVPR.2018.00884)
85. Sutton RS. 1990 Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In *Proc. of the Seventh Int. Conf. on Machine Learning*, pp. 216–224.
86. Sutton RS, Barto AG. 2018 *Reinforcement learning, second edition: an introduction*. Cambridge, MA: MIT Press.
87. Sutton RS, Barto AG. 1981 An adaptive network that constructs and uses and internal model of its world. *Cogn. Brain Theory* **4**, 217–246.
88. Craik KJW. 1943 *The nature of explanation*. Cambridge, UK: Cambridge University Press.
89. Sutton RS. 1988 Learning to predict by the methods of temporal differences. *Machine Learn.* **3**, 9–44. (doi:10.1007/BF00115009)
90. Watkins CJCH. 1989 Learning from delayed rewards. PhD, University of Cambridge, UK. See https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.330022
91. Watkins CJCH, Dayan P. 1992 Q-learning. *Mach. Learn.* **8**, 279–292. (doi:10.1007/BF00992698)
92. Mnih V *et al.* 2015 Human-level control through deep reinforcement learning. *Nature* **518**, Article 7540. (doi:10.1038/nature14236)
93. Silver D *et al.* 2016 Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, Article 7587. (doi:10.1038/nature16961)
94. Tesauro G. 1995 Temporal difference learning and TD-Gammon. *Commun. ACM* **38**, 58–68. (doi:10.1145/203330.203343)
95. Kumaran D, Hassabis D. 2015 From Pixels to Actions: Human-level control through Deep Reinforcement Learning. *Google AI Blog*. See http://ai.googleblog.com/2015/02/from-pixels-to-actions-human-level.html
96. Levine S, Finn C, Darrell T, Abbeel P. 2016 End-to-End Training of Deep Visuomotor Policies. *ArXiv:1504.00702 [Cs]*. See http://arxiv.org/abs/1504.00702
97. Zhu Y, Mottaghi R, Kolve E, Lim JJ, Gupta A, Fei-Fei L, Farhadi A. 2016 Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning. *ArXiv:1609.05143 [Cs]*. See http://arxiv.org/abs/1609.05143
98. Mirowski P *et al*. 2017 Learning to Navigate in Complex Environments. *ArXiv:1611.03673 [Cs]*. See http://arxiv.org/abs/1611.03673
99. Mirowski P *et al*. 2019 Learning to Navigate in Cities Without a Map. *ArXiv:1804.00168 [Cs]*. See http://arxiv.org/abs/1804.00168
100. Lange S, Riedmiller M. 2010 Deep auto-encoder neural networks in reinforcement learning. In *The 2010 Int. Joint Conf. on Neural Networks (IJCNN)*, pp. 1–8. (doi:10.1109/IJCNN.2010.5596468)
101. Pan X, You Y, Wang Z, Lu C. 2017 Virtual to Real Reinforcement Learning for Autonomous Driving. *ArXiv:1704.03952 [Cs]*. See http://arxiv.org/abs/1704.03952
102. Codevilla F, Müller M, López A, Koltun V, Dosovitskiy A. 2018 End-to-end Driving via Conditional Imitation Learning. *ArXiv:1710.02410 [Cs]*. See http://arxiv.org/abs/1710.02410
103. Dosovitskiy A, Koltun V. 2017 Learning to Act by Predicting the Future. *ArXiv:1611.01779 [Cs]*. See http://arxiv.org/abs/1611.01779
104. Levine S, Shah D. 2022 Learning robotic navigation from experience: principles, methods, and recent results. *Phil. Trans. R. Soc. B* **377**, 20210447. (doi:10.1098/rstb.2021.0447)
105. Zhou B, Krähenbühl P, Koltun V. 2019 Does computer vision matter for action? *Sci. Robot.* **4**, eaaw6661. (doi:10.1126/scirobotics.aaw6661)
106. Eslami SMA *et al.* 2018 Neural scene representation and rendering. *Science* **360**, 1204–1210. (doi:10.1126/science.aar6170)
107. Sax A, Emi B, Zamir AR, Guibas L, Savarese S, Malik J. 2019 Mid-Level Visual Representations Improve Generalization and Sample Efficiency for Learning Visuomotor Policies. *ArXiv:1812.11971 [Cs]*. See http://arxiv.org/abs/1812.11971
108. Mousavian A, Toshev A, Fiser M, Kosecka J, Wahid A, Davidson J. 2019 Visual Representations for Semantic Target Driven Navigation. *ArXiv:1805.06066 [Cs]*. See http://arxiv.org/abs/1805.06066
109. Chen B, Sax A, Lewis G, Armeni I, Savarese S, Zamir A, Malik J, Pinto L. 2020 Robust Policies via Mid-Level Visual Representations: An Experimental Study in Manipulation and Navigation. *ArXiv:2011.06698 [Cs]*. See http://arxiv.org/abs/2011.06698
110. Ravichandran Z, Peng L, Hughes N, Griffith JD, Carlone L. 2021 Hierarchical Representations and Explicit Memory: Learning Effective Navigation Policies on 3D Scene Graphs using Graph Neural Networks. *ArXiv:2108.01176 [Cs]*. See http://arxiv.org/abs/2108.01176
111. Rosano M, Furnari A, Gulino L, Santoro C, Farinella GM. 2022 Image-based Navigation in Real-World Environments via Multiple Mid-level Representations: Fusion Models, Benchmark and Efficient Evaluation. *ArXiv:2202.01069 [Cs]*. See http://arxiv.org/abs/2202.01069
112. Muryy A, Siddharth N, Nardelli N, Torr PHS, Glennerster A. 2020 Lessons from reinforcement learning for biological representations of space. *ArXiv:1912.06615 [q-Bio]*. See http://arxiv.org/abs/1912.06615
113. Laflaquière A, O'Regan JK, Argentieri S, Gas B, Terekhov AV. 2015 Learning agent's spatial configuration from sensorimotor invariants. *Robot. Auton. Syst.* **71**, 49–59. (doi:1016/j.robot.2015.01.003)
114. Laflaquière A, O'Regan JK, Gas B, Terekhov A. 2018 Discovering space—Grounding spatial topology and metric regularity in a naive agent's sensorimotor experience. *Neural Netw.* **105**, 371–392. (doi:10.1016/j.neunet.2018.06.001)
115. Laflaquière A, Garcia Ortiz M. 2019 Unsupervised emergence of egocentric spatial structure from sensorimotor prediction. *Adv. Neural Inf. Process. Syst.* **32**, 1–11.
116. Laflaquière A. 2020 Emergence of Spatial Coordinates via Exploration. *ArXiv Preprint ArXiv:2010.15469*.
117. Papert SA. 1966 The Summer Vision Project. *AI Memos*. See https://dspace.mit.edu/handle/1721.1/6125
118. Choy CB, Xu D, Gwak J, Chen K, Savarese S. 2016 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. *ArXiv:1604.00449 [Cs]*. See http://arxiv.org/abs/1604.00449
119. Girdhar R, Fouhey DF, Rodriguez M, Gupta A. 2016 Learning a Predictable and Generative Vector Representation for Objects. *ArXiv:1603.08637 [Cs]. See* http://arxiv.org/abs/1603.08637
120. Wu J, Xue T, Lim JJ, Tian Y, Tenenbaum JB, Torralba A, Freeman WT. 2016 Single Image 3D Interpreter

Network. *ArXiv:1604.08685 [Cs]*, 9910, 365–382. (doi:10.1007/978-3-319-46466-4_22)

121. Tatarchenko M, Dosovitskiy A, Brox T. 2016 Multi-view 3D Models from Single Images with a Convolutional Network. *ArXiv:1511.06702 [Cs]*. See http://arxiv.org/abs/1511.06702
122. Jimenez Rezende D, Eslami SM, Mohamed S, Battaglia P, Jaderberg M, Heess N. 2016 Unsupervised learning of 3D structure from images. *Adv. Neural Inf. Process. Syst.* **29**, 1–9. (doi:10.48550/arXiv.1607.00662)
123. Kato H, Ushiku Y, Harada T. 2017 Neural 3D Mesh Renderer. *ArXiv:1711.07566 [Cs]*. See http://arxiv.org/abs/1711.07566
124. Tatarchenko M, Richter SR, Ranftl R, Li Z, Koltun V, Brox T. 2019 What Do Single-view 3D Reconstruction Networks Learn? *ArXiv:1905.03678 [Cs]*. See http://arxiv.org/abs/1905.03678
125. Groueix T, Fisher M, Kim VG, Russell BC, Aubry M. 2018 AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. *ArXiv:1802.05384 [Cs]*. See http://arxiv.org/abs/1802.05384
126. Richter SR, Roth S. 2018 Matryoshka Networks: Predicting 3D Geometry via Nested Shape Layers. *ArXiv:1804.10975 [Cs]*. See http://arxiv.org/abs/1804.10975
127. Tatarchenko M, Dosovitskiy A, Brox T. 2017 Octree Generating Networks: Efficient Convolutional Architectures for High-resolution 3D Outputs. In *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2107–2115. (doi:10.1109/ICCV.2017.230)
128. Saxena A, Sun M, Ng AY. 2007 *Learning* 3-D Scene Structure from a Single Still Image. In *2007 IEEE 11th Int. Conf. on Computer Vision*, pp 1–8. (doi:10.1109/ICCV.2007.4408828)
129. Malik J *et al.* 2016 The three R's of computer vision. *Pattern Recognit. Lett.* **72**, 4–14. (doi:10.1016/j.patrec.2016.01.019)
130. Song S, Yu F, Zeng A, Chang AX, Savva M, Funkhouser T. 2016 Semantic Scene Completion from a Single Depth Image. *ArXiv:1611.08974 [Cs]*. See http://arxiv.org/abs/1611.08974
131. Rosenbaum D, Besse F, Viola F, Rezende DJ, Eslami SMA. 2018 Learning models for visual 3D localization with implicit mapping. *ArXiv:1807.03149 [Cs, Stat]*. See http://arxiv.org/abs/1807.03149
132. Ha D, Schmidhuber J. 2018 Recurrent World Models Facilitate Policy Evolution. *ArXiv:1809.01999 [Cs, Stat]*. See http://arxiv.org/abs/1809.01999
133. Sitzmann V, Zollhöfer M, Wetzstein G. 2019 Scene representation networks: continuous 3D-structure-aware neural scene representations. *Adv. Neural Inf. Process. Syst.* **32**, 1–12. (doi:10.48550/arXiv.1906.01618)
134. Xie Y *et al*. 2022 Neural Fields in Visual Computing and Beyond. *ArXiv:2111.11426 [Cs]*. See http://arxiv.org/abs/2111.11426
135. Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A. 2019 Occupancy Networks: Learning 3D Reconstruction in Function Space. *ArXiv:1812.03828 [Cs]*. See http://arxiv.org/abs/1812.03828
136. Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S. 2019 DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. *ArXiv:1901.05103 [Cs]*. See http://arxiv.org/abs/1901.05103
137. Niemeyer M, Mescheder L, Oechsle M, Geiger A. 2020 Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. *ArXiv:1912.07372 [Cs, Eess]*. See http://arxiv.org/abs/1912.07372
138. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. 2020 NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *ArXiv:2003.08934 [Cs]*. See http://arxiv.org/abs/2003.08934
139. Yu A, Fridovich-Keil S, Tancik M, Chen Q, Recht B, Kanazawa A. 2021 Plenoxels: Radiance Fields without Neural Networks. *ArXiv:2112.05131 [Cs]*. See http://arxiv.org/abs/2112.05131
140. Yu A, Ye V, Tancik M, Kanazawa A. 2021 pixelNeRF: Neural Radiance Fields from One or Few Images. *ArXiv:2012.02190 [Cs]*. See http://arxiv.org/abs/2012.02190
141. Kosiorek AR, Strathmann H, Zoran D, Moreno P, Schneider R, Mokrá S, Rezende DJ. 2021 *NeRF-VAE: A Geometry Aware 3D Scene Generative Model* (arXiv:2104.00587). arXiv. See (doi:10.48550/arXiv.2104.00587)
142. Sitzmann V, Rezchikov S, Freeman WT, Tenenbaum J. B, Durand F. 2021 Light Field Networks: Neural Scene Representations with Single-Evaluation Rendering. *ArXiv:2106.02634 [Cs]*. See http://arxiv.org/abs/2106.02634
143. Adelson EH, Bergen JR. 1991 The Plenoptic Function and the Elements of Early Vision. In *Computational models of visual processing* (eds M. Landy, J. A. Movshon), pp. 3–20. Cambridge, MA: MIT Press.
144. Gibson JJ. 1966 *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
145. Nagata T *et al.* 2012 Depth perception from image defocus in a jumping spider. *Science* **335**, 469–471. (doi:10.1126/science.1211667)
146. Guo Q, Shi Z, Huang Y-W, Alexander E, Qiu C-W, Capasso F, Zickler T. 2019 Compact single-shot metalens depth sensors inspired by eyes of jumping spiders. *Proc. Natl Acad. Sci. USA* **116**, 22 959–22 965. (doi:10.1073/pnas.1912154116)
147. Chung W-S, Marshall J. 2014 Range-finding in squid using retinal deformation and image blur. *Curr. Biol.* **24**, R64–R65. (doi:10.1016/j.cub.2013.11.058)
148. Fan Q *et al.* 2022 Trilobite-inspired neural nanophotonic light-field camera with extreme depth-of-field. *Nat. Commun.* **13**, Article 1. (doi:10.1038/s41467-022-29568-y)
149. Banks M, Sprague W, Schmoll J, Parnell J, Love G. 2015 Why do animal eyes have pupils of different shapes? *Sci. Adv.* **1**, e1500391. (doi:10.1126/sciadv.1500391)
150. Harkness L. 1977 Chameleons use accommodation cues to judge distance. *Nature* **267**, Article 5609. (doi:10.1038/267346a0)
151. Nityananda V, Read JCA. 2017 Stereopsis in animals: evolution, function and mechanisms. *J. Exp. Biol.* **220**, 2502–2512. (doi:10.1242/jeb.143883)
152. Nityananda V, Tarawneh G, Henriksen S, Umeton D, Simmons A, Read JCA. 2018 A novel form of stereo vision in the praying mantis. *Curr. Biol.* **28**, 588–593.e4. (doi:10.1016/j.cub.2018.01.012)
153. Read JCA. 2021 Binocular vision and stereopsis across the animal kingdom. *Annu. Rev. Vis. Sci.* **7**, 389–415. (doi:10.1146/annurev-vision-093019-113212)
154. Morgan M. 2022 Stereopsis for rapidly moving targets. *Phil. Trans. R. Soc. B* **377**, 20210462. (doi:10.1098/rstb.2021.0462)
155. Read J. 2022 Stereopsis without correspondence. *Phil. Trans. R. Soc. B* **377**, 20210449. (doi:10.1098/rstb.2021.0449)
156. van der Willigen RF, Frost BJ, Wagner H. 1998 Stereoscopic depth perception in the owl. *Neuroreport* **9**, 1233–1237. (doi:10.1097/00001756-199804200-00050)
157. Collett T. 1977 Stereopsis in toads. *Nature* **267**, 349. (doi:10.1038/267349a0)
158. Feord RC, Sumner ME, Pusdekar S, Kalra L, Gonzalez-Bellido PT, Wardill TJ. 2020 Cuttlefish use stereopsis to strike at prey. *Sci. Adv.* **6**, eaay6036. (doi:10.1126/sciadv.aay6036)
159. Jordanous A. 2020 Intelligence without representation: a historical perspective. *Systems* **8**, Article 3. (doi:10.3390/systems8030031)
160. Webb B. 2000 What does robotics offer animal behaviour? *Anim. Behav.* **60**, 545–558. (doi:10.1006/anbe.2000.1514)
161. Webb B. 2001 Can robots make good models of biological behaviour? *Behav. Brain Sci.* **24**, 1033–1050. (doi:10.1017/S0140525X01000127)
162. Webb B. 2002 Robots in invertebrate neuroscience. *Nature* **417**, Article 6886. (doi:10.1038/417359a)
163. Webb B. 2020 Robots with insect brains. *Science* **368**, 244–245. (doi:10.1126/science.aaz6869)
164. Helgadottir L, Haenicke J, Landgraf T, Rojas R, Nawrot M. 2013 Conditioned behavior in a robot controlled by a spiking neural network. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 891–894. (doi:10.1109/NER.2013.6696078)
165. Stone T *et al.* 2017 An anatomically constrained model for path integration in the bee brain. *Curr. Biol.* **27**, 3069–3085.e11. (doi:10.1016/j.cub.2017.08.052)
166. Marshall J. 2021 Borrowing bee brains. *New Scientist* **249**, 23. (doi:10.1016/S0262-4079(21)00279-7)
167. Chancán M, Hernandez-Nunez L, Narendra A, Barron AB, Milford M. 2020 A hybrid compact neural architecture for visual place recognition. *IEEE Robot. Autom. Lett.* **5**, 993–1000. (doi:10.1109/LRA.2020.2967324)
168. Kamhi JF, Barron AB, Narendra A. 2020 Vertical lobes of the mushroom bodies are essential for view-based navigation in Australian Myrmecia Ants. *Curr. Biol.* **30**, 3432–3437.e3. (doi:10.1016/j.cub.2020.06.030)

169. Buehlmann C, Wozniak B, Goulard R, Webb B, Graham P, Niven JE. 2020 Mushroom bodies are required for learned visual navigation, but not for innate visual behavior, in ants. *Curr. Biol.* **30**, 3438–3443.e2. (doi:10.1016/j.cub.2020.07.013)
170. Heinze S. 2020 Visual navigation: ants lose track without mushroom bodies. *Curr. Biol.* **30**, R984–R986. (doi:10.1016/j.cub.2020.07.038)
171. Tolman EC. 1948 Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208. (doi:10.1037/h0061626)
172. Marozzi E, Jeffery KJ. 2012 Place, space and memory cells. *Curr. Biol.* **22**, R939–R942. (doi:10.1016/j.cub.2012.10.022)
173. O'Keefe J, Dostrovsky J. 1971 The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175. (doi:10.1016/0006-8993(71)90358-1)
174. O'Keefe J, Nadel L. 1978 *The hippocampus as a cognitive Map*. Oxford, UK: Clarendon Press.
175. O'Keefe J, Nadel L. 1979 Précis of O'Keefe & Nadel's *The hippocampus as a cognitive map*. *Behav. Brain Sci.* **2**, 487–494. (doi:10.1017/S0140525X00063949)
176. Fyhn M, Molden S, Witter MP, Moser EI, Moser M-B. 2004 Spatial representation in the entorhinal cortex. *Science* **305**, 1258–1264. (doi:10.1126/science.1099901)
177. Hafting T, Fyhn M, Molden S, Moser M-B, Moser EI. 2005 Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, Article 7052. (doi:10.1038/nature03721)
178. Moser EI, Moser M-B. 2008 A metric for space. *Hippocampus* **18**, 1142–1156. (doi:10.1002/hipo.20483)
179. Bush D, Barry C, Manson D, Burgess N. 2015 Using grid cells for navigation. *Neuron* **87**, 507–520. (doi:10.1016/j.neuron.2015.07.006)
180. Grieves RM, Jeffery KJ. 2017 The representation of space in the brain. *Behav. Processes* **135**, 113–131. (doi:10.1016/j.beproc.2016.12.012)
181. Rowland DC, Roudi Y, Moser M-B, Moser EI. 2016 Ten years of grid cells. *Annu. Rev. Neurosci.* **39**, 19–40. (doi:10.1146/annurev-neuro-070815-013824)
182. Carpenter F, Barry C. 2016 Distorted grids as a spatial label and metric. *Trends Cogn. Sci.* **20**, 164–167. (doi:10.1016/j.tics.2015.12.004)
183. Barry C, Hayman R, Burgess N, Jeffery KJ. 2007 Experience-dependent rescaling of entorhinal grids. *Nat. Neurosci.* **10**, 6. (doi:10.1038/nn1905)
184. Stensola H, Stensola T, Solstad T, Frøland K, Moser M-B, Moser EI. 2012 The entorhinal grid map is discretized. *Nature* **492**, 7427. (doi:10.1038/nature11649)
185. O'Keefe J, Burgess N. 1996 Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**, 6581. (doi:10.1038/381425a0)
186. Jeffery KJ. 2015 Distorting the metric fabric of the cognitive map. *Trends Cogn. Sci.* **19**, 300–301. (doi:10.1016/j.tics.2015.04.001)
187. Bellmund JLS, de Cothi W, Ruiter TA, Nau M, Barry C, Doeller CF. 2020 Deforming the metric of cognitive maps distorts memory. *Nat. Hum. Behav.* **4**, 2. (doi:10.1038/s41562-019-0767-). 3.
188. Stensola T, Stensola H, Moser M-B, Moser EI. 2015 Shearing-induced asymmetry in entorhinal grid cells. *Nature* **518**, 7538. (doi:10.1038/nature14151)
189. Krupic J, Bauza M, Burton S, Barry C, O'Keefe J. 2015 Grid cell symmetry is shaped by environmental geometry. *Nature* **518**, 7538. (doi:10.1038/nature14153)
190. Krupic J, Bauza M, Burton S, O'Keefe J. 2018 Local transformations of the hippocampal cognitive map. *Science* **359**, 1143–1146. (doi:10.1126/science.aao4960)
191. Boccara CN, Nardin M, Stella F, O'Neill J, Csicsvari J. 2019 The entorhinal cognitive map is attracted to goals. *Science* **363**, 1443–1447. (doi:10.1126/science.aav4837)
192. Butler WN, Hardcastle K, Giocomo LM. 2019 Remembered reward locations restructure entorhinal spatial maps. *Science* **363**, 1447–1452. (doi:10.1126/science.aav5297)
193. Jeffery KJ, Jovalekic A, Verriotis M, Hayman R. 2013 Navigating in a three-dimensional world. *Behav. Brain Sci.* **36**, 523–543. (doi:10.1017/S0140525X12002476)
194. Horiuchi TK, Moss CF. 2015 Grid cells in 3-D: reconciling data and models. *Hippocampus* **25**, 1489–1500. (doi:10.1002/hipo.22469)
195. Stella F, Treves A. 2015 The self-organization of grid cells in 3D. *ELife* **4**, e05913. (doi:10.7554/eLife.05913)
196. Mathis A, Stemmler MB, Herz AV. 2015 Probable nature of higher-dimensional symmetries underlying mammalian grid-cell activity patterns. *ELife* **4**, e05979. (doi:10.7554/eLife.05979)
197. Finkelstein A, Las L, Ulanovsky N. 2016 3-D maps and compasses in the brain. *Annu. Rev. Neurosci.* **39**, 171–196. (doi:10.1146/annurev-neuro-070815-013831)
198. Aronov D. 2021 Disordered grids in the third dimension. *Nat. Neurosci.* **24**, 11. (doi:10.1038/s41593-021-00925-2)
199. Jeffery KJ. 2021 How environmental movement constraints shape the neural code for space. *Cogn. Process* **22**, 97–104. (doi:10.1007/s10339-021-01045-2)
200. Jeffery K. 2022 Symmetries and asymmetries in the neural encoding of 3D space. *Phil. Trans. R. Soc. B* **377**, 20210452. (doi:10.1098/rstb.2021.0452)
201. Momennejad I. 2022 A rubric for human-like and neuroAI. *Phil. Trans. R. Soc. B* **377**, 20210446. (doi:10.1098/rstb.2021.0446)
202. Brooks R, Hassabis D, Bray D, Shashua A. 2012 Is the brain a good model for machine intelligence? *Nature* **482**, Article 7386. (doi:10.1038/482462a)
203. Hassabis D, Kumaran D, Summerfield C, Botvinick M. 2017 Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258. (doi:10.1016/j.neuron.2017.06.011)
204. Richards BA *et al.* 2019 A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, Article 11. (doi:10.1038/s41593-019-0520-2)
205. Zador A *et al*. 2022 *Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution* (arXiv:2210.08340). arXiv. See (doi:10.48550/arXiv.2210.08340)
206. Mataric MJ. 1991 Navigating with a rat brain: A neurobiologically-inspired model for robot spatial representation. In *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats*, pp. 169–175.
207. Burgess N, O'Keefe J, Burgess N, Donnett JG, Jeffery KJ, O-keefe J. 1997 Robotic and neuronal simulation of the hippocampus and rat navigation. *Phil. Trans. R. Soc. Lond. B* **352**, 1535–1543. (doi:10.1098/rstb.1997.0140)
208. Arleo A. 2000 Spatial learning and navigation in neuro-mimetic systems: modeling the rat hippocampus, PhD thesis, EPFL. (https://www.epfl.ch/en/)
209. Arleo A, Gerstner W. 2000 Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biol. Cybern* **83**, 287–299. (doi:10.1007/s004220000171)
210. Milford MJ, Wyeth GF, Prasser D. 2004 RatSLAM: A hippocampal model for simultaneous localization and mapping. In *IEEE Int. Conf. on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, 1*, vol. **1**, pp. 403–408. (doi:10.1109/ROBOT.2004.1307183)
211. Milford MJ. 2007 *Robot navigation from nature: simultaneous localisation, mapping, and path planning based on hippocampal models*. Berlin, Germany: Springer.
212. Milford MJ, Wyeth GF. 2008 Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Trans. Rob.* **24**, 1038–1053. (doi:10.1109/TRO.2008.2004520)
213. Milford M, Wyeth G. 2010 Hybrid robot control and SLAM for persistent navigation and mapping. *Robot. Auton. Syst.* **58**, 1096–1104. (doi:10.1016/j.robot.2010.05.004)
214. Milford MJ, Wiles J, Wyeth GF. 2010 Solving navigational uncertainty using grid cells on robots. *PLoS Comput. Biol.* **6**, e1000995. (doi:10.1371/journal.pcbi.1000995)
215. Nyberg N, Duvelle É, Barry C, Spiers HJ. 2022 Spatial goal coding in the hippocampal formation. *Neuron* **110**, 394–422. (doi:10.1016/j.neuron.2021.12.012)
216. Banino A *et al.* 2018 Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 7705. (doi:10.1038/s41586-018-0102-6)
217. Cueva CJ, Wei X-X 2018 *Emergence of grid-like representations by training recurrent neural networks to perform spatial localization* (arXiv:1803.07770). arXiv. See (doi:10.48550/arXiv.1803.07770)
218. Fiete IR, Burak Y, Brookings T. 2008 What grid cells convey about rat location. *J. Neurosci.* **28**, 6858–6871. (doi:10.1523/JNEUROSCI.5684-07.2008)
219. Erdem UM, Hasselmo M. 2012 A goal-directed spatial navigation model using forward trajectory

planning based on grid cells. *Eur. J. Neurosci.* **35**, 916–931. (doi:10.1111/j.1460-9568.2012.08015.x)

220. Momennejad I. 2020 Learning structures: predictive representations, replay, and generalization. *Curr. Opin. Behav. Sci.* **32**, 155–166. (doi:10.1016/j.cobeha.2020.02.017)
221. Sarel A, Finkelstein A, Las L, Ulanovsky N. 2017 Vectorial representation of spatial goals in the hippocampus of bats. *Science* **355**, 176–180. (doi:10.1126/science.aak9589)
222. Howard LR, Javadi AH, Yu Y, Mill RD, Morrison LC, Knight R, Loftus MM, Staskute L, Spiers HJ. 2014 The hippocampus and entorhinal cortex encode the path and Euclidean distances to goals during navigation. *Curr. Biol.* **24**, 1331–1340. (doi:10.1016/j.cub.2014.05.001)
223. Gustafson NJ, Daw ND. 2011 Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Comput. Biol.* **7**, e1002235. (doi:10.1371/journal.pcbi.1002235)
224. Stachenfeld KL, Botvinick MM, Gershman SJ. 2014 Design principles of the hippocampal cognitive map. *Adv. Neural Inf. Process. Syst.* **3**, 2528–2536.
225. Stachenfeld KL, Botvinick MM, Gershman SJ. 2017 The hippocampus as a predictive map. *Nat. Neurosci.* **20**, Article 11. (doi:10.1038/nn.4650)
226. Dayan P. 1993 Improving generalization for temporal difference learning: the successor representation. *Neural Comput.*. **5**, 613–624. (doi:10.1162/neco.1993.5.4.613)
227. Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ. 2017 The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, Article 9. (doi:10.1038/s41562-017-0180-8)
228. Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. 2017 Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* **13**, e1005768. (doi:10.1371/journal.pcbi.1005768)
229. Geerts JP, Chersi F, Stachenfeld KL, Burgess N. 2020 A general model of hippocampal and dorsal striatal learning and decision making. *Proc. Natl Acad. Sci. USA* **117**, 31 427–31 437. (doi:10.1073/pnas.2007981117)
230. Solstad T, Moser EI, Einevoll GT. 2006 From grid cells to place cells: a mathematical model. *Hippocampus* **16**, 1026–1031. (doi:10.1002/hipo.20244)
231. Dordek Y, Soudry D, Meir R, Derdikman D. 2016 Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *ELife* **5**, e10094. (doi:10.7554/eLife.10094)
232. Sorscher B, Mel G, Ganguli S, Ocko S. 2019 A unified theory for the origin of grid cells through the lens of pattern formation. In *Advances in Neural Information Processing Systems, 32*. See https://proceedings.neurips.cc/paper/2019/hash/6e7d5d259be7bf56ed79029c4e621f44-Abstract.html
233. Saleem AB, Diamanti EM, Fournier J, Harris KD, Carandini M. 2018 Coherent encoding of subjective spatial position in visual cortex and hippocampus. *Nature* **562**, 124–127. (doi:10.1038/s41586-018-0516-1)
234. Horrocks E, Mareschal I, Saleem A. 2022 Walking humans and running mice: perception and neural encoding of optic flow during self-motion. *Phil. Trans. R. Soc. B* **377**, 20210450. (doi:10.1098/rstb.2021.0450)
235. Gibson JJ. 1950 *The perception of the visual world*. Boston, MA: Houghton Mifflin Co.
236. Gibson JJ. 1979 *The ecological approach To visual perception*. New York, NY: Psychology Press.
237. Wagner M, Hatfield G, Cassese K, Makwinski AN. 2018 Differentiating between affine and perspective-based models for the geometry of visual space based on judgments of the interior angles of squares. *Vision* **2**, Article 2. (doi:10.3390/vision2020022)
238. Warren WH. 2012 Does This Computational Theory Solve the Right Problem? Marr, Gibson, and the Goal of Vision. *Perception* **41**, 1053–1060. (doi:10.1068/p7327)
239. Tsao T, Tsao DY. 2022 A topological solution to object segmentation and tracking. *Proc. Natl Acad. Sci. USA* **119**, e2204248119. (doi:10.1073/pnas.2204248119)
240. O'Regan JK, Noë A. 2001 A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* **24**, 939–973; discussion 973–1031.
241. Glennerster A. 2022 Understanding 3D vision as a policy network. *Phil. Trans. R. Soc. B* **377**, 20210448. (doi:10.1098/rstb.2021.0448)
242. Pizlo Z. 2010 *3D shape: Its unique place in visual perception*. Cambridge, MA: MIT Press.
243. Pizlo Z, Li Y, Sawada T. 2014 *Making a machine that sees like Us*. Oxford, UK: Oxford University Press.
244. Clark JJ, Yuille AL. 1990 *Data fusion for sensory information processing systems*. Berlin, Germany: Springer.
245. Knill DC, Richards W. 1996 *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.
246. Landy MS, Maloney LT, Johnston EB, Young M. 1995 Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res.* **35**, 389–412. (doi:10.1016/0042-6989(94)00176-m)
247. Trommershauser J, Kording K, Landy MS. 2011 *Sensory cue integration*. Oxford, UK: Oxford University Press.
248. Ma WJ, Kording K, Goldreich D. 2022 *Bayesian models of perception and action*. Cambridge, MA: MIT Press.
249. Landy MS, Banks MS, Knill DC. 2011 Ideal-Observer Models of Cue Integration. In *Sensory Cue integration* (eds J Trommershauser, K Kording, MS Landy), pp. 5–29. Oxford, UK: Oxford University Press.
250. Johnston EB, Cumming BG, Parker AJ. 1993 Integration of depth modules: stereopsis and texture. *Vision Res.* **33**, 813–826. (doi:10.1016/0042-6989(93)90200-G)
251. Johnston EB, Cumming BG, Landy MS. 1994 Integration of stereopsis and motion shape cues. *Vision Res.* **34**, 2259–2275. (doi:10.1016/0042-6989(94)90106-6)
252. Jacobs RA. 1999 Optimal integration of texture and motion cues to depth. *Vision Res.* **39**, 3621–3629. (doi:10.1016/s0042-6989(99)00088-7)
253. Hillis JM, Ernst MO, Banks MS, Landy MS. 2002 Combining sensory information: mandatory fusion within, but not between, senses. *Science* **298**, 1627–1630. (doi:10.1126/science.1075396)
254. Knill DC, Saunders JA. 2003 Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Res.*, **43**(24), 2539–2558. (doi:10.1016/S0042-6989(03)00458-9)
255. Hillis JM, Watt SJ, Landy MS, Banks MS. 2004 Slant from texture and disparity cues: optimal cue combination. *J. Vis.* **4**, 1. (doi:10.1167/4.12.1)
256. Watt SJ, Akeley K, Ernst MO, Banks MS. 2005 Focus cues affect perceived depth. *J. Vis.* **5**, 834–862. (doi:10.1167/5.10.7)
257. Glennerster A, Tcheang L, Gilson SJ, Fitzgibbon AW, Parker AJ. 2006 Humans ignore motion and stereo cues in favor of a fictional stable world. *Curr. Biol.* **16**, 428–432. (doi:10.1016/j.cub.2006.01.019)
258. Girshick AR, Banks MS. 2009 Probabilistic combination of slant information: weighted averaging and robustness as optimal percepts. *J. Vis.* **9**, 8. (doi:10.1167/9.9.8)
259. Svarverud E, Gilson SJ, Glennerster A. 2010 Cue combination for 3D location judgements. *J. Vis.* **10**, 5. (doi:10.1167/10.1.5)
260. Scarfe P, Hibbard PB. 2011 Statistically optimal integration of biased sensory estimates. *J. Vis.* **11**, 12–12. (doi:10.1167/11.7.12)
261. Lovell PG, Bloj M, Harris JM. 2012 Optimal integration of shading and binocular disparity for depth perception. *J. Vis.* **12**, 1–18. (doi:10.1167/12.1.1)
262. Domini F, Caudek C. 2011 Combining image signals before three-dimensional reconstruction: the intrinsic constraint model of cue integration. In *Sensory Cue integration* (eds J Trommershäuser, K Kording, MS Landy), pp. 120–143. Oxford, UK: Oxford University Press. (doi:10.1093/acprof:oso/9780195387247.003.0007)
263. Landy MS, Kojima H. 2001 Ideal cue combination for localizing texture-defined edges. *J. Opt. Soc. Am., Opt. Image Sci. Vis.* **18**, 2307–2320. (doi:10.1364/josaa.18.002307)
264. Oruç I, Maloney L, Landy M. 2003 Weighted linear cue combination with possibly correlated error. *Vision Res.* **43**, 2451–2468. (doi:10.1016/S0042-6989(03)00435-8)
265. Rosas P, Wichmann FA, Wagemans J. 2007 Texture and object motion in slant discrimination: failure of reliability-based weighting of cues may be evidence for strong fusion. *J. Vis.* **7**, 3. (doi:10.1167/7.6.3)
266. Foster R, Fantoni C, Caudek C, Domini F. 2011 Integration of disparity and velocity information for haptic and perceptual judgments of object depth. *Acta Psychologica* **136**, 300–310. (doi:10.1016/j.actpsy.2010.12.003)
267. Chen C-C, Tyler CW. 2015 Shading beats binocular disparity in depth from luminance gradients: evidence against a maximum likelihood principle for

cue combination. *PLoS ONE* **10**, e0132658. (doi:10.1371/journal.pone.0132658)

268. Rahnev D, Denison RN. 2018 Suboptimality in perceptual decision making. *Behav. Brain Sci.* **41**, e223. (doi:10.1017/S0140525X18000936)
269. Rosas P, Wichmann FA. 2011 Cue combination: Beyond optimality. In *Sensory Cue integration* (eds J Trommershauser, K Kording, MS Landy), pp. 144–152. Oxford, UK: Oxford University Press.
270. Scarfe P. 2022 Experimentally disambiguating models of sensory cue integration. *J. Vis.* **22**, 5. (doi:10.1167/jov.22.1.5)
271. Bülthoff HH, Mallot HA. 1990 Integration of Stereo, Shading, and Texture. In *AI and the eye* (eds A Blake, T Troscianko), pp. 119–146. New York, NY: Wiley.
272. Tyler CW. 2020 An accelerated cue combination principle accounts for multi-cue depth perception. *J. Percept. Imaging* **3**, 10501-1. (doi:10.2352/j.percept.imaging.2020.3.1.010501)
273. Yuille AL, Bülthoff HH. 1996 Bayesian decision theory and psychophysics. In *Perception as Bayesian inference* (eds DC Knill, W Richards), pp. 123–162. Cambridge, UK: Cambridge University Press.
274. Kersten D. 1997 Inverse 3-D graphics: a metaphor for visual perception. *Behav. Res. Methods Instrum. Comput.* **29**, 37–46. (doi:10.3758/BF03200564)
275. Kersten D, Yuille A. 2003 Bayesian models of object perception. *Curr. Opin Neurobiol.* **13**, 150–158. (doi:10.1016/S0959-4388(03)00042-4)
276. Kersten D, Mamassian P, Yuille A. 2004 Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304. (doi:10.1146/annurev.psych.55.090902.142005)
277. Yuille A, Kersten D. 2006 Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308. (doi:10.1016/j.tics.2006.05.002)
278. Battaglia PW, Hamrick JB, Tenenbaum JB. 2013 Simulation as an engine of physical scene understanding. *Proc. Natl Acad. Sci. USA* **110**, 18 327–18 332. (doi:10.1073/pnas.1306572110)
279. Ullman TD, Spelke E, Battaglia P, Tenenbaum JB. 2017 Mind Games: game engines as an architecture for intuitive physics. *Trends Cogn. Sci.* **21**, 649–665. (doi:10.1016/j.tics.2017.05.012)
280. Erdogan G, Jacobs RA. 2017 Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychol. Rev.* **124**, 740–761. (doi:10.1037/rev0000086)
281. Yildirim I, Belledonne M, Freiwald W, Tenenbaum J. 2020 Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, eaax5979. (doi:10.1126/sciadv.aax5979)
282. Fleming RW. 2014 Visual perception of materials and their properties. *Vision Res.* **94**, 62–75. (doi:10.1016/j.visres.2013.11.004)
283. Fleming RW. 2017 Material perception. *Annu. Rev. Vision Sci.* **3**, 365–388. (doi:10.1146/annurev-vision-102016-061429)
284. Fleming RW, Storrs KR. 2019 Learning to see stuff. *Curr. Opin. Behav. Sci.* **30**, 100–108. (doi:10.1016/j.cobeha.2019.07.004)
285. DiCarlo JJ *et al*. 2021 *How does the brain combine generative models and direct discriminative computations in high-level vision?* See https://openreview.net/forum?id=zlTiwFtLlR4
286. Geisler WS. 2011 Contributions of ideal observer theory to vision research. *Vision Res.* **51**, 771–781. (doi:10.1016/j.visres.2010.09.027)
287. Johnston EB. 1991 Systematic distortions of shape from stereopsis. *Vision Res.* **31**, 1351–1360. (doi:10.1016/0042-6989(91)90056-B)
288. Todd JT, Norman JF. 2003 The visual perception of 3-D shape from multiple cues: are observers capable of perceiving metric structure? *Percept. Psychophys.* **65**, 31–47. (doi:10.3758/BF03194781)
289. Campagnoli C, Hung B, Domini F. 2022 Explicit and implicit depth-cue integration: evidence of systematic biases with real objects. *Vision Res.* **190**, 107961. (doi:10.1016/j.visres.2021.107961)
290. Vishwanath D. 2014 Toward a new theory of stereopsis. *Psychol. Rev.* **121**, 151–178. (doi:10.1037/a0035233)
291. Peacocke C. 1983 *Sense and content: experience, thought and their relations*. Oxford, UK: Oxford University Press.
292. Hibbard PB. 2008 Can appearance be so deceptive? Representationalism and binocular vision. *Spat. Vis.* **21**, 549–559. (doi:10.1163/156856808786451444)
293. Hornsey RL, Hibbard PB, Scarfe P. 2015 Ordinal judgments of depth in monocularly- and stereoscopically-viewed photographs of complex natural scenes. In *2015 International Conference on 3D Imaging (IC3D)*, pp. 1–5. (doi:10.1109/IC3D.2015.7391812)
294. Di Luca M, Domini F, Caudek C. 2010 Inconsistency of perceived 3D shape. *Vision Res.* **50**, 1519–1531. (doi:10.1016/j.visres.2010.05.006)
295. Koenderink JJ. 1998 Pictorial relief. *Phil. Trans. R. Soc. A* **356**, 1071–1086. (doi:10.1098/rsta.1998.0211)
296. Loomis JM, Da Silva JA, Fujita N, Fukusima SS. 1992 Visual space perception and visually directed action. *J. Exp. Psychol. Hum. Percept. Perf.* **18**, 906–921. (doi:10.1037//0096-1523.18.4.906)
297. Loomis JM, Philbeck JW, Zahorik P. 2002 Dissociation between location and shape in visual space. *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 1202–1212. (doi:10.1037/0096-1523.28.5.1202)
298. Loomis JM. 2014 Three theories for reconciling the linearity of egocentric distance perception with distortion of shape on the ground plane. *Psychol. Neurosci.* **7**, 245–251. (doi:10.3922/j.psns.2014.034)
299. Koenderink JJ, van Doorn AJ, Kappers AML, Lappin JS. 2002 Large-scale visual frontoparallels under full-cue conditions. *Perception* **31**, 1467–1475. (doi:10.1068/p3295)
300. Svarverud E, Gilson S, Glennerster A. 2012 A demonstration of 'broken' visual space. *PLoS ONE* **7**, e33782. (doi:10.1371/journal.pone.0033782)
301. Gillam B, Chambers D. 1985 Size and position are incongruous: measurements on the Müller-Lyer figure. *Percept. Psychophys.* **37**, 549–556. (doi:10.3758/BF03204921)
302. Smeets JBJ, Sousa R, Brenner E. 2009 Illusions can warp visual space. *Perception* **38**, 1467–1480. (doi:10.1068/p6439)
303. Morales J, Bax A, Firestone C. 2020 Sustained representation of perspectival shape. *Proc. Natl Acad. Sci. USA* **117**, 14 873–14 882. (doi:10.1073/pnas.2000715117)
304. Morales J, Bax A, Firestone C. 2021 Reply to Linton: perspectival interference up close. *Proc. Natl Acad. Sci. USA* **118**, e2025440118. (doi:10.1073/pnas.2025440118)
305. Morales J, Firestone C. In press. Empirical evidence for perspectival similarity. *Psych. Rev.*
306. Linton P. 2021 Conflicting shape percepts explained by perception cognition distinction. *Proc. Natl Acad. Sci. USA* **118**, e2024195118. (doi:10.1073/pnas.2024195118)
307. Burge J, Burge T. 2022 Shape, perspective, and what is and is not perceived: comment on Morales, Bax, and Firestone (2020). *Psychol. Rev.* (doi:10.1037/rev0000363)
308. Vishwanath D. 2010 Reconciling Pictures and Reality: Information in Surface and Depth Perception. In *Perception beyond inference: The information content of visual processes* (eds L Albertazzi, GJ van Tonder, D Vishwanath), pp. 201–240. Cambridge, MA: MIT Press.
309. Linton P. 2017 *The perception and cognition of visual space*. Cham, Switzerland: Palgrave Macmillan.
310. Linton P. 2022 Minimal theory of 3D vision: new approach to visual scale and visual shape. *Phil. Trans. R. Soc. B* **378**, 20210455. (doi:10.1098/rstb.2021.0455)
311. Kepler J. 1604 Paralipomena to Witelo. In *Optics: paralipomena to witelo and optical part of astronomy* (ed. WH Donahue (Trans.)), 2000. Santa Fe, NM: Green Lion Press.
312. Descartes R. 1637 *Dioptrique (Optics)*. In *The philosophical writings of descartes: volume 1 (1985)* (eds J Cottingham, R Stoothoff, D Murdoch), pp. 152–175. Cambridge, UK: Cambridge University Press.
313. Gillam B. 1998 Illusions at Century's End. In *Perception and cognition at century's End* (ed. J Hochberg), pp. 95–136. Cambridge, MA: Academic Press.
314. Koenderink J, van Doorn A, Wagemans J. 2015 Part and whole in pictorial relief. *I-Perception* **6**, 2041669515615713. (doi:10.1177/2041669515615713)
315. Horn B, Brooks MJ. 1989 *Shape from shading*. Cambridge, MA: MIT Press.
316. Koenderink J, Doorn Av, Albertazzi L, Wagemans J. 2015 Relief articulation techniques. *Art Percept.* **3**, 151–171. (doi:10.1163/22134913-00002032)
317. Zucker SW. 2000 On qualitative shape inferences: a journey from geometry to topology. arXiv:2008.08622v1. (doi:10.48550/arXiv.2008.08622)

318. Kunsberg B, Zucker SW. 2018 Critical contours: an invariant linking image flow with salient surface organization. *SIAM J. Imaging Sci.* **11**, 1849–1877. (doi:10.1137/17M1145525)
319. Kunsberg B, Zucker SW. 2021 From boundaries to bumps: when closed (extremal) contours are critical. *J. Vis.* **21**, 7. (doi:10.1167/jov.21.13.7)
320. Koenderink JJ, van Doorn AJ, Wagemans J. 2018 Geometry of pictorial relief. *Annu. Rev. Vision Sci.* **4**, 451–474. (doi:10.1146/annurev-vision-091517-034250)
321. Domini F, Braunstein ML. 1998 Recovery of 3-D structure from motion is neither Euclidean nor affine. *J. Exp. Psychol.: Hum. Percept. Perform.* **24**, 1273–1295. (doi:10.1037/0096-1523.24.4.1273)
322. Tittle JS, Todd JT, Perotti VJ, Norman JF. 1995 Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *J. Exp. Psychol.: Hum. Percept. Perform.* **21**, 663–678. (doi:10.1037/0096-1523.21.3.663)
323. Todd JT, Tittle JS, Norman JF. 1995 Distortions of three-dimensional space in the perceptual analysis of motion and stereo. *Perception* **24**, 75–86. (doi:10.1068/p240075)
324. Todd JT, Oomes AHJ, Koenderink JJ, Kappers AML. 2001 On the affine structure of perceptual space. *Psychol. Sci.* **12**, 191–196. (doi:10.1111/1467-9280.00335)
325. Wagner M, Gambino AJ. 2016 Variations in the anisotropy and affine structure of visual space: a geometry of visibles with a third dimension. *Topoi* **35**, 583–598. (doi:10.1007/s11245-015-9303-x)
326. Glennerster A, Vuong J, Fitzgibbon A. 2015 Modelling observers' errors in pointing to an unseen target. *J. Vis.* **15**, 1080. (doi:10.1167/15.12.1080)
327. Domini F, Caudek C. 2013 Perception and action without veridical metric reconstruction: an affine approach. In *Shape perception in human and computer vision*, pp. 285–298. London, UK: Springer.
328. Domini F. 2022 A deterministic theory of visual 3D processing. *Phil. Trans. R. Soc. B* **377**, 20210458. (doi:10.1098/rstb.2021.0458)
329. Gilinsky AS. 1951 Perceived size and distance in visual space. *Psychol. Rev.* **58**, 460–482. (doi:10.1098/rstb.2021.0458)
330. Hatfield G. 2003 Representation and constraints: the inverse problem and the structure of visual space. *Acta Psychologica* **114**, 355–378. (doi:10.1016/j.actpsy.2003.07.003)
331. Hatfield G. 2012 Phenomenal and cognitive factors in spatial perception. In *Visual experience: sensation, cognition, and constancy* (eds G Hatfield, S Allred), pp. 35–62. Oxford, UK: Oxford University Press.
332. Erkelens CJ. 2015 The extent of visual space inferred from perspective angles. *I-Perception* **6**, 5–14. (doi:10.1068/i0673)
333. Erkelens CJ. 2021 Geometric constraints of visual space. *I-Perception* **12**, 20416695211055212. (doi:10.1177/20416695211055212)
334. Vishwanath D. 2022 Modelling the phenomenology and psychophysics of 3D perception from pictures to reality. *Phil. Trans. R. Soc. B* **377**, 20210454. (doi:10.1098/rstb.2021.0454)
335. Engel A, Maye A, Kurthen M, König P. 2013 Where's the action? The pragmatic turn in cognitive science. *Trends Cogn. Sci.* **17**, 202–209. (doi:10.1016/j.tics.2013.03.006)
336. Engel AK, Friston KJ, Kragic D. 2016 *The pragmatic turn: toward action-oriented views in cognitive science*. Cambridge, MA: MIT Press.
337. Creem-Regehr SH, Kunz BR. 2010 Perception and action. *WIREs Cogn. Sci.* **1**, 800–810. (doi:10.1002/wcs.82)
338. Warren WH. 2019 Perceiving Surface Layout: Ground Theory, Affordances, and the Objects of Perception. In *Perception as information detection*, pp. 151–173. Abingdon, UK: Routledge.
339. Creem-Regehr S, Stefanucci J, Bodenheimer R. 2022 Perceiving distance in virtual reality: theoretical insights from contemporary technologies. *Phil. Trans. R. Soc. B* **378**, 20210456. (doi:10.1098/rstb.2021.0456)
340. Levine S, Shah D. 2022 Learning robotic navigation from experience: principles, methods, and recent results. *Phil. Trans. R. Soc. B* **378**, 20210447. (doi:10.1098/rstb.2021.0447)
341. Gibson JJ. 1979 The ecological approach to visual perception. New York, NY: Psychology Press.
342. Warren WH. 2019 Perceiving surface layout: ground theory, affordances, and the objects of perception. In *Perception as information detection*, pp. 608–703. Abingdon, UK: Routledge. (doi:10.1037//0096-1523.10.5.683)
343. Proffitt DR. 2006 Embodied perception and the economy of action. *Perspect. Psychol. Sci.* **1**, 110–122. (doi:10.1111/j.1745-6916.2006.00008.x)
344. Proffitt DR. 2009 Affordances matter in geographical slant perception. *Psychon. Bull. Rev.* **16**, 970–972. (doi:10.3758/PBR.16.5.970)
345. Bhalla M, Proffitt DR. 1999 Visual-motor recalibration in geographical slant perception. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 1076–1096. (doi:10.1037//0096-1523.25.4.1076)
346. Witt JK, Linkenauger SA, Bakdash JZ, Proffitt DR. 2008 Putting to a bigger hole: golf performance relates to perceived size. *Psychon. Bull. Rev.* **15**, 581–585. (doi:10.3758/PBR.15.3.581)
347. Stefanucci JK, Proffitt DR, Clore GL, Parekh N. 2008 Skating down a steeper slope: fear influences the perception of geographical slant. *Perception* **37**, 321–323. (doi:10.1068/p5796)
348. Witt JK. 2011 Action's effect on perception. *Curr. Direct. Psychol. Sci.* **20**, 201–206. (doi:10.1177/0963721411408770)
349. Durgin FH, Baird JA, Greenburg M, Russell R, Shaughnessy K, Waymouth S. 2009 Who is being deceived? The experimental demands of wearing a backpack. *Psychon. Bull. Rev.* **16**, 964–969. (doi:10.3758/PBR.16.5.964)
350. Firestone C. 2013 How 'paternalistic' is spatial perception? Why wearing a heavy backpack doesn't—and couldn't—make hills look steeper. *Perspect. Psychol. Sci.* **8**, 455–45 . (doi:10.1177/1745691613489835)
351. Firestone C, Scholl BJ. 2016 Cognition does not affect perception: evaluating the evidence for 'top-down' effects. *Behav. Brain Sci.* **39**, e229. (doi:10.1017/S0140525X15000965)
352. Schnall S. 2017*a* Social and contextual constraints on embodied perception. *Perspect. Psychol. Sci.: J. Assoc. Psychol. Sci.* **12**, 325–340. (doi:10.1177/1745691616660199)
353. Firestone C, Scholl BJ. 2017 Seeing and thinking in studies of embodied 'perception.' *Perspect. Psychol. Sci.: J. Assoc. Psychol. Sci.* **12**, 341–343. (doi:10.1177/1745691616679944)
354. Durgin FH. 2017 Counterpoint: distinguishing between perception and judgment of spatial layout. *Perspect. Psychol. Sci.* **12**, 344–346. (doi:10.1177/1745691616677829)
355. Schnall S. 2017*b* No magic bullet in sight: a reply to Firestone and Scholl (2017) and Durgin (2017). *Perspect. Psychol. Sci.* **12**, 347–349. (doi:10.1177/1745691617691948)
356. Glennerster A, Rogers BJ, Bradshaw MF. 1996 Stereoscopic depth constancy depends on the subject's task. *Vision Res.* **36**, 3441–3456. (doi:10.1016/0042-6989(96)00090-9)
357. Bradshaw MF, Parton AD, Glennerster A. 2000 The task-dependent use of binocular disparity and motion parallax information. *Vision Res.* **40**, 3725–3734. (doi:10.1016/S0042-6989(00)00214-5)
358. Norman JF, Crabtree CE, Clayton AM, Norman HF. 2005 The perception of distances and spatial relationships in natural outdoor environments. *Perception* **34**, 1315–1324. (doi:10.1068/p5304)
359. Goodale MA, Milner AD. 1992 Separate visual pathways for perception and action. *Trends Neurosci.* **15**, 20–25. (doi:10.1016/0166-2236(92)90344-8)
360. Milner AD, Goodale MA. 1995 *The visual brain in action*. Oxford, UK: Oxford University Press.
361. Goodale M, Milner D. 2013 *Sight unseen: An exploration of conscious and unconscious vision*. Oxford, UK: Oxford University Press.
362. Whitwell R, Garach M, Goodale, M Sperandio, I. 2022 Looking at the Ebbinghaus illusion: differences in neurocomputational requirements, not gaze-mediated attention, explain a classic perception–action dissociation. *Phil. Trans. R. Soc. B* **378**, 20210459. (doi:10.1098/rstb.2021.0459)
363. Horrocks E, Mareschal I, Saleem A. 2022 Walking humans and running mice: perception and neural encoding of optic flow during self-motion. *Phil. Trans. R. Soc. B* **378**, 20210450. (doi:10.1098/rstb.2021.0450)
364. Maselli A, Ofek E, Cohn B, Hinckley K, Gonzalez-Franco M. 2022 Enhanced efficiency in visually guided online motor control for actions directed towards the body midline. *Phil. Trans. R. Soc. B* **378**, 20210453. (doi:10.1098/rstb.2021.0453)

365. Lanman DR. 2020 Display systems research at facebook reality labs (conference presentation). *Opt. Archit. Disp. Sens. Augment. Virtual Mix. Real.* **11310**, 1131025. (doi:10.1117/12.2566597)
366. Jaloza LB. 2022 *Passing the visual Turing test: The inside story of our quest for visual realism in VR*. Tech at Meta. See https://tech.fb.com/ar-vr/2022/06/passing-the-visual-turing-test-the-inside-story-of-our-quest-for-visual-realism-in-vr/
367. Guan P, Mercier O, Shvartsman M, Lanman D. 2022 Perceptual Requirements for Eye-Tracked Distortion Correction in VR. In *ACM SIGGRAPH 2022 Conf, Proc.*, 1–8. See (doi:10.1145/3528233.3530699)
368. Tong J, Allison RS, Wilcox LM, Allison RS, Wilcox LM. 2019 The Impact of Radial Distortions in VR Headsets on Perceived Surface Slant. *J. Imaging Sci. Technol.* **63**, 1–11. (doi:10.2352/J.ImagingSci.Technol.2019.63.6.060409)
369. Tong J, Allison RS, Wilcox LM. 2020 Optical distortions in VR bias the perceived slant of moving surfaces. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 73–79. (doi:10.1109/ISMAR50242.2020.00027)
370. Rzepka AM, Hussey KJ, Maltz MV, Babin KP, Wilcox L, Culham J. 2022 Familiar size affects perception differently in virtual reality and the real world. *Phil. Trans. R. Soc. B* **378**, 20210464. (doi:10.1098/rstb.2021.0464)
371. Hess RF, To L, Zhou J, Wang G, Cooperstock JR. 2015 Stereo vision: the haves and have-nots. *I-Perception* **6**, 2041669515593028. (doi:10.1177/2041669515593028)
372. Levi DM. 2022 Learning to see in depth. *Vision Res.* **200**, 108082. (doi:10.1016/j.visres.2022.108082)
373. O'Connor AR, Birch EE, Anderson S, Draper H, the FSOS Research Group. 2010 The Functional Significance of Stereopsis. *Invest. Ophthalmol. Vis. Sci.*, **51**, 2019–2023. doi:10.1167/iovs.09-4434
374. Melmoth DR, Grant S. 2006 Advantages of binocular vision for the control of reaching and grasping. *Exp. Brain Res.* **171**, 371–388. (doi:10.1007/s00221-005-0273-x)
375. Bonnen K, Matthis JS, Gibaldi A, Banks MS, Levi DM, Hayhoe M. 2021 Binocular vision and the control of foot placement during walking in natural terrain. *Sci. Rep.* **11**, Article 1. (doi:10.1038/s41598-021-99846-0)
376. Barry SR. 2009 *Fixing My gaze: a scientist's journey into seeing in three dimensions*. New York, NY: Basic Books.
377. Sacks O. 2010 *The mind's eye*. New York, NY: Knopf.
378. Bridgeman B. 2014 Restoring adult stereopsis: a vision researcher's personal experience. *Optom. Vis. Sci.* **91**, e135. (doi:10.1097/OPX.00000000 00000272)
379. Niechwiej-Szwedo E, Colpa L, Wong A. 2022 The role of binocular vision in the control and development of visually guided upper limb movements. *Phil. Trans. R. Soc. B* **378**, 20210461. (doi:10.1098/rstb.2021.0461)
380. Fine I, Wade AR, Brewer AA, May MG, Goodman DF, Boynton GM, Wandell BA, MacLeod DIA. 2003 Long-term deprivation affects visual perception and cortex. *Nat. Neurosci.* **6**, 915–916. (doi:10.1038/nn1102)
381. Huber E, Webster J, Brewer A, MacLeod D, Wandell B, Wade A, Fine I. 2015 A lack of experience-dependent plasticity after more than a decade of recovered sight. *Psychol. Sci.* **26**, 393–401. (doi:10.1177/0956797614563957)
382. Fine I, Park WJ. 2022 Do you hear what I see? How do early blind individuals experience object motion? *Phil. Trans. R. Soc. B* **378**, 20210460. (doi:10.1098/rstb.2021.0460)
383. Glennerster A (Director). 2015 *Andrew Glennerster's lab at the University of Reading*. See https://www.youtube.com/watch?v=7TXMQ2_0NAE
384. Wolpert D. 2017 *Probabilistic models of sensorimotor control. Cognitive Computational Neuroscience.* See https://www.youtube.com/watch?v=wTYHF4LAKQI)
385. Muthukrishna M, Henrich J. 2019 A problem in theory. *Nat. Hum. Behav.* **3**, Article 3. (doi:10.1038/s41562-018-0522-1)
386. Oberauer K, Lewandowsky S. 2019 Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* **26**, 1596–1618. (doi:10.3758/s13423-019-01645-2)
387. Fried EI. 2020 Lack of theory building and testing impedes progress in the factor and network literature. *Psychol. Inq.* **31**, 271–288. (doi:10.1080/1047840X.2020.1853461)
388. Wolpert D. 2021 *How the brain controls the body*. 2021 Ferrier Lecture, Royal Society. See https://www.youtube.com/watch?v=WGtcD2UDUyA
389. Nurse P. 2021 Biology must generate ideas as well as data. *Nature* **597**, 305–305. (doi:10.1038/d41586-021-02480-z)
390. Chu JSG, Evans JA. 2021 Slowed canonical progress in large fields of science. *Proc. Natl Acad. Sci. USA* **118**, e2021636118. (doi:10.1073/pnas.2021636118)
391. Eronen MI, Bringmann LF. 2021 The theory crisis in psychology: how to move forward. *Perspect. Psychol. Sci.* **16**, 779–788. (doi:10.1177/1745691620970586)
392. Burghardt J, Bodansky AN. 2021 Why psychology needs to stop striving for novelty and how to move towards theory-driven research. *Front. Psychol.* **12**, 1–10. (doi:10.3389/fpsyg.2021.609802)
393. van Rooij I, Baggio G. 2021 Theory before the test: how to build high-verisimilitude explanatory theories in psychological science. *Perspect. Psychol. Sci.* **16**, 682–697. (doi:10.1177/1745691620970604)
394. McPhetres J *et al.* 2021 A decade of theory as reflected in Psychological Science (2009–2019). *PLoS ONE* **16**, e0247986. (doi:10.1371/journal.pone.0247986)
395. Borsboom D, van der Maas HLJ, Dalege J, Kievit RA, Haig BD. 2021 Theory construction methodology: a practical framework for building theories in psychology. *Perspect. Psychol. Sci.* **16**, 756–766. (doi:10.1177/1745691620969647)
396. Robinaugh DJ, Haslbeck JMB, Ryan O, Fried EI, Waldorp LJ. 2021 Invisible hands and fine calipers: a call to use formal theory as a toolkit for theory construction. *Perspect. Psychol. Sci.* **16**, 725–743. (doi:10.1177/1745691620974697)
397. Maatman FO. 2021 *Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It*. PsyArXiv. See (doi:10.31234/osf.io/puqvs)
398. Press C, Yon D, Heyes C. 2022 Building better theories. *Curr. Biol.* **32**, R13–R17. (doi:10.1016/j.cub.2021.11.027)