Predictive coding in the brain and deep neural networks

Zhaoyang Pang

▶ **To cite this version:**

Zhaoyang Pang. Predictive coding in the brain and deep neural networks. Neurons and Cognition [q-bio.NC]. Université Paul Sabatier - Toulouse III, 2022. English. NNT : 2022TOU30037 . tel-03714369

HAL Id: tel-03714369
https://theses.hal.science/tel-03714369

Submitted on 5 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Zhaoyang PANG

Le 28 février 2022

Codage prédictif dans le cerveau et les réseaux de neurones profonds

Ecole doctorale : **CLESCO - Comportement, Langage, Education, Socialisation, Cognition**

Spécialité : **Neurosciences**

Unité de recherche :
CERCO - Centre de Recherche Cerveau et Cognition

Thèse dirigée par
Rufin VANRULLEN

Jury

M. Laurent PERRINET, Rapporteur
Mme Laura DUGUE, Examinatrice
M. Matthew CHALK, Examinateur
Mme Mathilde BONNEFOND, Examinatrice
M. Rufin VANRULLEN, Directeur de thèse
M. Jean MARTINET, Président

Acknowledgments

Now, my Ph.D life is finished. For the doctoral work I got involved in, there are still a lot of things to improve such as theoretical thinking and necessary research skills; for myself, I feel proud that I can finish the challenge. Yes, it's a challenge, a big one. The reason that I feel proud is that I did not give up or stop trying. However, from a professional view, my work is far from satisfactory. Why? one technical reason is that I did not accumulate the required knowledge before and during the Ph.D. Another big reason is that I did not have the experience or successful experience to deal with the situation where I was scared by the difficulties. But after going through anxiety, fear, and struggle, I started to think about how to become a real brave human. I think that is one of the valuable experience I got from the Ph.D. Being brave is easy because one can choose to behave bravely in the next second, but being brave is also difficult because a brave person can also behave cowardly suddenly. I want to thank my supervisor, thesis jury members, colleagues, and friends because they support me and encourage me, and most important show me how brave they are.

First I would like to thank my supervisor: Rufin VanRullen. Thank you for the constant and regular feedback on the research work. I believe without your help and professional guidance, it would be impossible for me to progress and improve. I started knowing nothing about this field and you brought me to the research world. Thank you so much. Aside from that, I also want to thank you for showing me the good qualities of you as a scientist. You showed me how a well-organized person you are. I thought before most successful scientists only focused on their work but you not only devote yourself to the research work but also enjoy your life and hobbies. It got me thinking about how to live my own life in the research world.

I would like to thank my reviewers Jean Martinet and Laurent Perrinet. Thanks for accepting and reviewing the thesis manuscript. Your simulating questions and suggestions made me rethink my work and gain more ways of understanding the same problem. I would also like to thank Laura Dugue, Matthew Chalk, and Mathilde Bonnefond. Thanks for the questions in terms of my work during the defense.

I would like to thank Andrea Alamia, Bhavin Choksi, and Canhuang Luo. You had been helping me in the past. Without you, it would be very difficult for me to solve problems in my life in France as a foreign student and also problems with the research. Thank you. I also want to thank Callum Biggs O'May, thanks for your patience, whenever I came to you with questions, you always explained them to me until I can understand. Thank you.

I would like to thank my colleagues Romain Bielawski, Mathieu Chalvidal, Colin Decourt, Benjamin Devillers, Milad Mozafari, Sabine Muzellec, Furkan Ozcelik, Victor Tinbou, Aimen Zerroug. You are smart guys. we had very nice conversations and I can always learn new stuff from you guys. Thanks a lot.

I would like to thank Anais Servais, Ludovic Gardy, Asmae Helmahi, and Xuening Li. You are very good friends. Your support and encouragement gave me strength. Finally, I would like to thank all the faculty in Cerco, especially Simon Thorpe, Berry-Isabelle, Nathalie, Claire, Pier-Giorgio Zanone, and Mahuffa. Without you, your help, and the nice environment you create, it's impossible for me to solve the problems I faced.

Contents

Abstract

Predictive coding is a famous and influential theory in the field of neuroscience. It states that instead of passively receiving the external information and forming perception or decision, the brain holds a hierarchical internal modal that could actively interact with the external stimuli. In this hierarchy, each level could predict the activation of the lower level, with the lowest level representing the outside world; while the predicted level could compute the difference between real stimulation and the prediction and send it upward to update the model for better prediction in the future. It could provide explanations for a wide range of neurophysiological and psychological phenomena, which leads to the belief that predictive coding may serve as a unifying computational framework for brain functions including sensation, perception, memory, and so on.

However, the interpretation provided by predictive coding on brain function is not definite enough. More supportive evidence is needed to establish predictive coding as a unifying principle used by the brain. On one hand, researchers build computational models implementing predictive coding dynamics and check whether such a dynamic system could generate some similar results observed in the biological brain. On the other hand, efforts have been made to investigate the neural mechanisms of predictive processing in the brain. In the current thesis, we combined both sides, experimental and computational, attempting to provide more solid evidence for predictive coding as a unifying theory of brain function.

Specifically, we conducted three progressive and related studies. The first one focuses on the neural activities in the biological brain, in particular, the oscillatory traveling waves. Could cortical traveling waves underlie predictive processes in the brain? Our results suggest that the ascending traveling waves only appear with the presence of bottom-up driven visual stimuli and disappear when visual inputs are absent; the descending waves, although they receive some modulation from external visual input, are less affected. We explained the results in the predictive coding framework: oscillatory traveling waves might be a neural signature of predictive coding with forward waves carrying prediction errors and backward waves trans-

mitting prediction signals.

In the second project, we utilized the deep learning technique and constructed a neural network which could implement predictive coding dynamics. If the human brain functions according to such a dynamic principle, the predictive coding network should show certain human-like properties. We tested this hypothesis with illusory contour image stimuli. The study suggests the neural network driven by predictive coding dynamics possesses illusory perception, supporting the possibility that the same dynamics strategy, i.e. predictive coding, might be shared between the network and biological brain.

Based on the results of the first two studies: (i) cortical traveling waves reflecting the neural mechanisms of predictive coding in the biological brain; (ii) the predictive neural network displaying human-like performance, in the third study, we further update the same predictive neural network by adding biologically plausible time delays and constants between network layers in order to generate oscillations. The preliminary results show that the network could oscillate with biologically plausible time parameters. We expect that such an oscillatory neural network will produce more human-like results in terms of its signal unit activation pattern and final decision output.

In summary, the thesis states the possibility of predictive coding theory as a unifying framework for brain functions by combining the evidence from the biological brain and computational neural network.

Résumé

Le codage prédictif est une théorie célèbre et influente dans le domaine des neurosciences. Il indique qu'au lieu de recevoir passivement les informations externes pour former une perception ou une décision, le cerveau utilise un modèle interne hiérarchique qui pourrait interagir activement avec les stimuli externes. Dans cette hiérarchie, chaque niveau prédirait l'activation du niveau inférieur, le niveau le plus bas représentant le monde extérieur ; tandis que le niveau prédit pourrait calculer la différence entre la stimulation réelle et la prédiction et l'envoyer vers le haut pour mettre à jour le modèle pour une meilleure prédiction à l'avenir. Cela pourrait fournir des explications pour un large éventail de phénomènes neurophysiologiques et psychologiques, ce qui conduit à croire que le codage prédictif peut servir de cadre unificateur pour les fonctions cérébrales, notamment la sensation, la perception, la mémoire, etc.

Cependant, l'interprétation raisonnable fournie par le codage prédictif ne peut exclure d'autres théories possibles sur la fonction cérébrale. Des preuves supplémentaires sont nécessaires pour prouver que le codage prédictif est un principe unificateur utilisé par le cerveau. D'une part, les chercheurs construisent des modèles informatiques mettant en œuvre une dynamique de codage prédictive et vérifient si un tel système dynamique pourrait générer des résultats similaires observés dans le cerveau biologique. D'autre part, des efforts ont été déployés pour étudier les mécanismes neuronaux du traitement prédictif dans le cerveau. Dans la thèse actuelle, nous avons combiné les deux côtés, expérimental et informatique, en essayant de fournir des preuves plus solides pour le codage prédictif en tant que théorie unificatrice de la fonction cérébrale.

Plus précisément, nous avons mené trois études progressives et connexes. La première se concentre sur les activités neuronales dans le cerveau biologique, en particulier les ondes oscillatoires progressives ('oscillatory travelling waves'). Les ondes progressives corticales pourraient-elles sous-tendre les processus prédictifs dans le cerveau ? Nos résultats suggèrent que les ondes ascendantes n'apparaissent qu'en présence de stimuli visuels et disparaissent

lorsque les entrées visuelles sont absentes ; tandis que les ondes descendantes, bien qu'elles reçoivent une certaine modulation de l'entrée visuelle externe, sont moins affectées. Nous avons expliqué les résultats dans le cadre du codage prédictif : les ondes progressives oscillatoires pourraient être le mécanisme neuronal du codage prédictif avec des ondes avant portant des erreurs de prédiction et des ondes arrière transmettant des signaux de prédiction.

Dans le deuxième projet, nous avons utilisé la technique d'apprentissage profond et construit un réseau de neurones qui pourrait mettre en œuvre une dynamique de codage prédictif. Si le cerveau humain pouvait fonctionner selon un tel principe dynamique, le réseau de codage prédictif devrait montrer certaines performances humaines. Nous avons testé cette hypothèse avec des stimuli d'image de contour illusoires. L'étude suggère que le réseau neuronal piloté par la dynamique de codage prédictif possède une perception illusoire, indiquant la possibilité que la même stratégie dynamique, c'est-à-dire le codage prédictif, pourrait être partagée entre le réseau et le cerveau biologique.

Sur la base des résultats des deux premières études : (i) les ondes progressives corticales pourraient servir de mécanisme neuronal de codage prédictif dans le cerveau biologique ; (ii) le réseau de neurones prédictifs affiche une perception similaire à l'humain, dans la troisième étude, nous renforçons davantage la plausibilité biologique du même réseau de neurones prédictif en ajoutant des constantes de temps et des délais et des entre les couches du réseau afin de générer des oscillations. Les résultats préliminaires montrent que le réseau pourrait osciller avec des paramètres temporels biologiquement plausibles. Nous nous attendons à ce qu'un tel réseau de neurones oscillatoire produise des résultats plus humains en termes de profil d'activation d'unité de signal et de sortie de décision finale.

En résumé, la thèse énonce la possibilité d'une théorie du codage prédictif en tant que cadre unificateur pour les fonctions cérébrales en combinant les preuves du cerveau biologique et du réseau neuronal informatique.

Résumé Substantiel

1. Introduction

Le cerveau humain peut remplir diverses fonctions, telles que la sensation, la perception, l'attention, la mémoire, etc. Quels sont les mécanismes internes du cerveau ? Comme nous le savons, la structure détermine la fonction. Par conséquent, pour comprendre le calcul neuronal du cerveau, il faut en déterminer l'agencement ou la constitution, puis, sur cette base, explorer plus avant la manière dont chaque composant se comporte pour générer une activité collective.

1.1 Traitement hiérarchique dans le cerveau

Au niveau macro, le cerveau peut être divisé en trois parties : le tronc cérébral, le cervelet et le cerveau antérieur. Le cerveau antérieur est la plus grande partie, dont la couche externe est appelée cortex cérébral et joue un rôle important dans les fonctions cérébrales. D'un point de vue microcosmique, les unités structurelles et fonctionnelles les plus fondamentales du système nerveux sont les cellules des neurones. Les neurones peuvent recevoir, intégrer et envoyer des signaux cérébraux. En outre, l'une des caractéristiques du cortex cérébral est sa structure laminaire. Le néocortex, qui occupe environ 90% du cortex cérébral, est formé de six couches distinctes. Comment le cerveau organise-t-il ces parties en un système complet et complexe capable d'effectuer un traitement sensoriel, une cognition et un contrôle ?

Anatomiquement, les connexions cortico-corticales dans le cerveau ont été étudiées et les connexions présentent une configuration hiérarchique. Dans le système visuel, par exemple, sa structure hiérarchique est définie par les schémas laminaires des origines et des terminaisons des projections cortico-corticales. Selon les premières études anatomiques (Rockland and Pandya, 1979; Felleman and Van Essen, 1991), trois directions de projection ont pu être définies: (i) les projections vers l'avant, ce qui signifie que les projections proviennent principalement des couches corticales supragranulaires et se terminent dans la couche granulaire (couche IV) ; (ii) projections en retour, provenant principalement des

couches corticales infragranulaires profondes et ciblant les couches superficielles et profondes; (iii) projections latérales, provenant des couches corticales supérieures et profondes dans une proportion plus égale et se terminant dans les zones cibles à la manière d'une colonne dans toutes les couches.

La configuration hiérarchique est liée au traitement hiérarchique dans le cerveau. D'une part, le long de la hiérarchie, la gamme de traitement devient plus large et les caractéristiques plus complexes. Par exemple, une des caractéristiques du système visuel est la rétinotopie ou topologie. La rétinotopie est une propriété de traitement des cellules dans la hiérarchie du système visuel, depuis la rétine jusqu'au cortex visuel. Elle fait référence à la manière topologique de mettre en correspondance l'entrée visuelle dans le champ visuel avec chaque niveau du flux visuel. En d'autres termes, les informations voisines dans le champ visuel, par exemple deux taches dans une image, sont traitées par des cellules spatialement proches dans le cerveau. De nombreuses structures cérébrales situées le long du flux visuel sont organisées en cartes rétinotopiques. Cependant, en montant dans la hiérarchie visuelle, la cartographie rétinotopique devient complexe.

D'autre part, le traitement dans la hiérarchie peut être divisé en différents modules ou voies et le traitement dans les différentes voies peut être parallèle ou distribué. Par exemple, dans le système visuel, il a été bien établi qu'il existe deux flux visuels (ou deux modules) dans le traitement visuel: les flux 'quoi' et 'où'. Comme leur nom l'indique, les flux 'quoi' traitent les informations liées à la reconnaissance, tandis que les flux 'où' traitent les informations liées à la localisation. Les deux flux s'interconnectent au niveau de V1 et se séparent ensuite en deux voies distinctes: La voie ventrale pour le courant 'quoi' et la voie dorsale pour le courant 'où'. De plus, cette organisation distribuée existe également dans d'autres modules ou systèmes. Il a été signalé qu'il existe également deux flux de traitement dans le système auditif. Par conséquent, la modularité de la structure hiérarchique peut être un mode d'organisation commun.

1.2 La théorie sur le cerveau: le codage prédictif

Nous avons maintenant acquis les connaissances de base sur la structure du cerveau. Les régions du cerveau sont reliées entre elles de manière hiérarchique. De plus, la représentation neuronale interne présente également des propriétés hiérarchiques dont la complexité et l'abstraction augmentent avec la structure hiérarchique. Comment les différentes fonctions cérébrales sont-elles réalisées sur la base d'une telle structure? Un algorithme de calcul potentiel adopté par le cerveau peut être le codage prédictif.

Dans un modèle interne hiérarchique mettant en œuvre une stratégie de codage prédictif, chaque niveau tente de prédire la représentation du niveau inférieur ou la stimulation externe pour le niveau le plus bas par une connexion de rétroaction. Cette prédiction sera comparée à la vérité de terrain reçue et, par conséquent, les discordances ou les erreurs de prédiction seront calculées au niveau inférieur. Grâce à la connexion feedforward de ce modèle, les erreurs de prédiction sont envoyées vers le haut pour corriger l'activité du niveau supérieur afin d'obtenir de meilleures prédictions la fois suivante. Cette théorie met l'accent sur l'interaction entre les prédictions descendantes et les pulsions ascendantes plutôt que sur la simple réalité sensorielle du monde extérieur. De manière incroyable, la théorie indique que la fonction cérébrale pourrait être réalisée sur la base d'un simple processus computationnel: la minimisation de l'erreur de prédiction.

Bien que le codage prédictif ait une logique très concise, il est censé expliquer un large éventail de phénomènes biologiques. Par conséquent, il pourrait servir de théorie globale sur le fonctionnement de l'ensemble du cerveau. Le codage prédictif a fait ses preuves dans la modélisation de phénomènes neurophysiologiques de bas niveau, notamment l'inhibition centre-souris dans la rétine, l'arrêt final dans V1, etc. Friston (2005) a proposé que le codage prédictif puisse sous-tendre la génération de réponses cérébrales évoquées. Il a souligné que les réponses corticales peuvent être considérées comme une manifestation observable d'un processus dans lequel le cerveau tente de minimiser l'erreur prédictive ou l'énergie libre engendrée par un stimulus sensoriel. Par exemple, les réponses évoquées dues à l'incongruence entre les informations dans la RF classique et en dehors de la RF peuvent être considérées comme un échec de la suppression de l'erreur de prédiction. De même, les

grands ERP causés par le conflit entre le traitement global et le traitement local pourraient également être traités comme des signaux d'erreur de prédiction dans le cortex cérébral.

1.3 Predictive coding models

Le cerveau adopte-t-il ou non une méthode de codage prédictif? Une façon de l'examiner est de construire des modèles de codage prédictif et de comparer ces modèles au cerveau biologique. Ces dernières années, quelques algorithmes de codage prédictif ont été mis en œuvre avec des techniques d'apprentissage profond, dans le but d'améliorer la reconnaissance des objets. Par exemple, le plus ancien réseau de codage prédictif profond est peut-être le PredNet (Lotter et al., 2016) qui a été utilisé pour prédire la prochaine image d'une séquence vidéo. En termes d'apprentissage du modèle, ils ont employé un objectif non supervisé, qui s'est avéré plus proche de l'apprentissage humain dans la réalité que les méthodes d'apprentissage supervisé standard. En effet, la plupart du temps, l'apprentissage humain consiste à décrire avec précision ce qu'il voit, au lieu de se faire dire ce qu'il voit par un instructeur. Néanmoins, il convient de mentionner que la capacité de leur modèle à prédire la prochaine image dans le flux vidéo peut être principalement due au fait qu'il a été entraîné à prédire.

Un autre exemple est Predify. L'implémentation de Predify par (Choksi et al., 2021) présente certaines similitudes avec les PCNs (Wen et al., 2018; Han et al., 2018). Les deux réseaux ont été conçus dans un but similaire d'amélioration de la reconnaissance des objets. Les différences sont également évidentes. En termes d'apprentissage du modèle ou de mise à jour des poids, Wen et al. (2018) a effectué l'optimisation uniquement avec un objectif de classification, ce qui entraîne une réduction non uniforme des erreurs de Reconstruction au cours des pas de temps. Cependant, le modèle Predify a été optimisé avec les deux objectifs de classification et de reconstruction, ce qui entraîne une performance de plus en plus meilleure au fil des pas de temps, ce qui est plus plausible sur le plan biologique. Dans la présente thèse, nous avons adapté le modèle décrit par (Choksi et al., 2021).

En résumé, bien que ces modèles puissent différer en termes de tâches, de fonctions

objectives et de méthodes d'apprentissage, leurs résultats montrent que le codage prédictif peut contribuer à améliorer les performances du modèle. Cela suggère qu'un modèle profond avec codage prédictif peut être plus performant que d'autres modèles ou même avoir le potentiel de montrer une performance semblable à celle d'un humain si nous concevons et implémentons soigneusement le modèle.

1.4 Preuve du codage prédictif dans le cerveau

Si le cerveau adopte effectivement une stratégie computationnelle de codage prédictif, quelle devrait être sa signature ou son reflet dans le cerveau. Les oscillations cérébrales constituent un candidat potentiel. Les oscillations cérébrales ou ondes cérébrales désignent l'activité neuronale rythmique dans le cortex cérébral. Elles ont été observées pour la première fois par Berger (1929) grâce à l'électroencéphalographie (EEG), qui est une méthode d'enregistrement de l'électrogramme des potentiels électriques du cuir chevelu. Dans son article fondateur, Berger (1929) a décrit deux signaux oscillatoires distincts qu'il a appelés rythme alpha et rythme bêta. Le premier, une oscillation relativement lente, signifiant une activité rythmique lente, se présentait à l'arrière de la tête des sujets avec les yeux fermés; le second, une oscillation relativement rapide apparaissait avec les yeux ouverts.

Les oscillations sont omniprésentes dans le cerveau et elles sont les composantes les plus importantes de la dynamique cérébrale. Initialement, les oscillations cérébrales sont considérées comme des sous-produits, comme le supposent les modèles neurophysiologiques conventionnels (Shadlen and Newsome, 1998). Mais il est maintenant évident que les fonctions des oscillations cérébrales sont étendues. Leurs fonctions peuvent être résumées grossièrement à deux niveaux: (i) Au niveau neuronal, comment les oscillations impliquent et coordonnent les activités neuronales; (ii) Au niveau cognitif, comment les oscillations influencent ou modulent diverses fonctions cognitives telles que l'attention, la conscience et la perception.

Les oscillations cérébrales peuvent fournir une évolution temporelle très fine, qui est étroitement liée à divers processus cognitifs. Dans le domaine spatial, un nombre croissant

d'études suggère que ces oscillations pourraient être organisées comme des ondes progressives à travers les régions du cerveau. Il est intéressant de noter que, dans le cadre du codage prédictif, la transmission des signaux (erreurs de prédiction et prédiction) s'effectue entre différents niveaux hiérarchiques, c'est-à-dire les régions du cerveau. Est-il possible que les ondes progressives ayant une certaine directionalité puissent transmettre les signaux postulés par la théorie du codage prédictif?

La mise en œuvre à grande échelle du codage prédictif dans le cerveau pourrait davantage impliquer les fréquences lentes, telles que les oscillations alpha. Alamia and VanRullen (2019) suggèrent que la propagation des oscillations alpha pourrait servir de mécanisme neuronal du codage prédictif à grande échelle. Ils ont d'abord construit un modèle à deux couches qui met en œuvre la dynamique du codage prédictif. Les résultats montrent que le modèle peut générer des oscillations de la bande alpha avec une constante de temps et un délai biologiquement plausibles. Dans leur deuxième expérience, ils ont élargi le modèle à plusieurs couches. Comme prévu, le modèle élargi pouvait générer des oscillations alpha se propageant à travers les couches du modèle. Il est remarquable que chaque couche du modèle élargi puisse correspondre à de grandes régions du cerveau, de la zone occipitale à la zone frontale, ce qui permet de comparer le modèle aux données empiriques obtenues dans ces zones. Les résultats informatiques et expérimentaux montrent que l'entrée feedforward entraîne des ondes alpha qui se déplacent des régions occipitales aux régions frontales, tandis que les signaux de rétroaction provoquent des ondes alpha qui passent dans la direction opposée. Autrement dit, à grande échelle, les signaux de prédiction pourraient être véhiculés par des ondes alpha descendantes, tandis que les erreurs de prédiction feedforward sont transmises par des ondes alpha ascendantes.

1.5 Les questions

Comme le titre le suggère, la présente thèse prend deux aspects-dans les cerveaux physiques et les réseaux de neurones profonds - pour évaluer la possibilité du codage prédictif comme théorie unificatrice du fonctionnement du cerveau. Du point de vue des cerveaux, le traitement prédictif dans les cerveaux, tel que postulé par la théorie, peut expliquer un large

éventail de phénomènes neurophysiologiques et psychologiques, comme examiné dans la section 1.3.2. Si la dynamique dans les cerveaux physiques est conduite par une stratégie de codage prédictif, une question naturelle se pose alors. Quel est le mécanisme neuronal sous-jacent? Ou encore, quels activités ou facteurs neuronaux peuvent entreprendre la tâche de transmission de l'information en termes d'erreurs de prédiction ascendantes et de signaux de prédiction descendants dans la hiérarchie corticale? Dans la section 1.5, j'ai examiné si les oscillations corticales, y compris les ondes progressives, pouvaient servir de mécanisme sous-jacent au codage prédictif, car elles peuvent jouer un rôle dans un large éventail de fonctions cérébrales, comme indiqué dans les sections 1.5.2 et 1.5.3. Certains chercheurs pensent que les oscillations gamma rapides pourraient transmettre les erreurs de prédiction dans la hiérarchie corticale, tandis que les oscillations alpha, beaucoup plus lentes, pourraient délivrer des signaux de prédiction descendants. Des opinions différentes existent en Alamia and VanRullen (2019) montrant que les deux signaux pourraient être transmis par l'oscillation alpha voyageant entre les zones corticales. Ainsi, il n'est toujours pas clair si les oscillations corticales pourraient sous-tendre le codage prédictif et quelles bandes de fréquences sont impliquées.

Un autre aspect concerne l'examen de la théorie des réseaux neuronaux profonds. Par rapport aux études empiriques, les approches computationnelles ne peuvent fournir que des preuves indirectes. Toutefois, on pense que la combinaison des deux méthodes peut apporter des éléments révélateurs. Les faits biologiques inspirent la construction de réseaux neuronaux artificiels ; à leur tour, les performances des réseaux neuronaux qui en résultent permettent de mieux comprendre la dynamique du cerveau. Revenons aux travaux actuels. Pour tirer parti des modèles de calcul alimentés par les techniques d'apprentissage profond (voir section 1.4.1), l'idée centrale est de construire un modèle de calcul avec une dynamique de codage prédictif et de s'attendre à ce qu'un tel réseau puisse afficher des performances similaires à celles du cerveau humain s'ils partagent le même système dynamique, c'est-à-dire le codage prédictif. Avant d'être transformée en modèle, la théorie doit être formulée correctement. La section 1.3.3 a présenté plusieurs façons possibles de formuler la même théorie - le codage prédictif. Le processus de construction et de mise en œuvre peut également varier.

La section 1.4.2 a passé en revue quelques réseaux neuronaux profonds différents pilotés par la dynamique du codage prédictif. Comme on peut le constater, de multiples facteurs, dont la manière de formuler une théorie et la manière de construire et d'entraîner un modèle, peuvent avoir une grande importance pour juger si la théorie, malgré son idée originale, peut fournir des prédictions précises.

Notre objectif final est de répondre à la question de savoir si le codage prédictif pourrait agir comme une théorie de la fonction cérébrale, qui peut être approchée en trouvant ses mécanismes neuronaux sous-jacents dans le cerveau. Pour concrétiser davantage le problème, nous pouvons nous demander si les ondes progressives oscillatoires peuvent servir de mécanisme potentiel, comme le suggère la section 1.5.4. Cependant, même si nous pouvons montrer les ondes progressives dans une certaine bande de fréquences qui peuvent présenter des schémas d'activation similaires à ceux suggérés par la théorie du codage prédictif, cela ne peut indiquer qu'une relation de corrélation et non de causalité. Ainsi, ce résultat possible doit être corroboré ailleurs. La méthode du modèle computationnel ou du réseau neuronal profond peut le prouver dans une direction opposée. C'est-à-dire que nous pouvons construire un réseau neuronal profond avec une dynamique de codage prédictive. Si un tel réseau peut montrer des oscillations ou des ondes progressives avec des paramètres biologiquement plausibles, nous pouvons être plus confiants dans notre hypothèse. Nous pouvons même aller plus loin en montrant que les oscillations ou les ondes progressives biologiques et artificielles peuvent affecter une certaine perception visuelle, par exemple l'illusion visuelle, de la même manière, puisque, en tant que modèle de vision, le réseau de neurones profonds peut traiter de multiples tâches visuelles, y compris la reconnaissance visuelle. Toutefois, avant d'aller aussi loin, il faudrait d'abord prouver que le réseau neuronal avec codage prédictif peut effectivement présenter des performances similaires à celles de l'homme, étant donné qu'il existe de multiples façons de formuler, de construire et d'entraîner un modèle.

Pour résumer, notre travail peut être divisé en trois études spécifiques: (i) pour trouver les mécanismes neuronaux pertinents du codage prédictif, nous nous demandons si les ondes

progressives oscillatoires peuvent servir d'événement neuronal pertinent possible en vérifiant leurs caractéristiques de traitement dans le cerveau; (ii) pour obtenir un réseau de codage prédictif de type humain, nous nous demandons si un tel réseau pourrait montrer une illusion visuelle de type humain; (iii) pour prouver davantage le rôle sous-jacent des oscillations, nous prenons le même réseau de codage prédictif et nous demandons s'il peut générer des oscillations biologiquement plausibles ou même des ondes progressives.

2. Les ondes alpha voyageuses comme corrélats neuronaux potentiels du codage prédictif ?

La première préoccupation de la thèse concerne l'implémentation neuronale du codage prédictif dans les cerveaux biologiques. Comme nous le savons déjà, dans le modèle hiérarchique hypothétique engagé dans le processus de codage prédictif, le niveau supérieur envoie des prédictions à un niveau inférieur, tandis que les signaux d'erreur sont transmis dans la direction opposée. Par conséquent, notre question concrète est de déterminer quelles activités neuronales dans le cerveau peuvent transporter les deux signaux bidirectionnels?

Les oscillations cérébrales peuvent servir de candidat possible. Cette spéculation peut provenir du fait que les oscillations et les ondes progressives ont un fort pouvoir d'explication pour un large éventail d'observations neurophysiologiques et psychophysiques, comme indiqué dans les sections 1.3.2 et 1.5.2. Plus important encore, une relation plus substantielle entre le codage prédictif et les oscillations peut exister si l'on considère la directionnalité de la transmission de leur message. Il a été démontré que les fréquences plus rapides (par exemple, gamma) pourraient transmettre un message vers l'avant, tandis que les fréquences plus basses (par exemple, alpha et bêta) sont liées à la transmission du message vers l'arrière : (Bastos et al., 2015a,b; Arnal and Giraud, 2012). Ces asymétries spectrales de l'oscillation cérébrale peuvent éclairer le passage du message dans un cadre de codage prédictif, les fréquences les plus rapides véhiculant les erreurs de prédiction ascendantes et les fréquences les plus basses les prédictions (Friston, 2019).

Par ailleurs, les ondes progressives pourraient même jouer un meilleur rôle dans le fonctionnement de la dynamique du codage prédictif en raison de leur directionnalité naturelle lorsqu'elles se propagent entre les régions du cerveau. Cette idée a été prouvée par une étude de modélisation (Alamia and VanRullen, 2019) où les oscillations alpha se déplaçaient vers le bas du modèle hiérarchique lorsque seuls des prieurs (prédictions) étaient proposés, tandis qu'une direction opposée des oscillations alpha se manifestait avec la seule présentation de l'entrée visuelle (signaux d'erreur). Il est remarquable que, dans l'architecture de codage prédictif, les oscillations alpha émergent naturellement avec des constantes de temps et des retards des neurones biologiquement plausibles. Cela indique un lien étroit entre les ondes alpha itinérantes et la réalisation du codage prédictif dans le cerveau biologique.

La première étude a été conçue pour fournir des preuves empiriques à l'étude de modélisation de Alamia and VanRullen (2019) pour enfin prouver le rôle des ondes progressives dans le codage prédictif. L'idée générale est de créer deux conditions de traitement distinctes, chacune supportant des ondes progressives avec une direction spécifique. Dans l'étude de modélisation, on peut faire passer des prédictions pures ou des signaux d'erreur à travers le modèle et observer les ondes progressives qui en résultent. Cependant, les choses pourraient être compliquées dans le cerveau, car celui-ci génère constamment des prédictions. Il serait possible de n'avoir que des prédictions dans le cerveau, par exemple en coupant l'entrée sensorielle ; cependant, nous ne pouvons pas obtenir une situation où seules les entrées feedforward sont impliquées en raison des prédictions existant en permanence. Par conséquent, dans la première étude, nous avons utilisé une méthode de quantification des ondes pour trier les ondes progressives et rétrospectives et nous avons analysé comment elles seraient liées à la transmission de messages dans un cadre de codage prédictif.

Sur la base des données EEG de participants humains, nous avons démontré que la direction d'une onde progressive (8-13 Hz) dépend de la tâche, confirmant les suggestions d'études antérieures (Zhang et al., 2018; Alamia and VanRullen, 2019; Halgren et al., 2019; Lozano-Soldevilla and VanRullen, 2019), et vérifiant les prédictions de notre propre modélisation étude sur la génération et la propagation des oscillations a (Alamia and

VanRullen, 2019). Plus précisément, nous avons caractérisé les ondes FW voyageant des régions occipitales aux régions pariétales suscitées par une stimulation visuelle, et les ondes BW dans la direction inversée dominant pendant l'état de repos. De plus, la présence d'une stimulation visuelle externe a réduit les ondes BW, ce qui est en accord avec d'autres études sur les ondes voyageuses spontanées (Patten et al., 2012; Sato et al., 2012). Enfin, pendant la stimulation visuelle, les ondes FW et les ondes BW étaient présentes et modulées par le type de stimulation (statique ou dynamique), mais elles étaient négativement corrélées dans le temps.

La génération et la directionnalité des ondes progressives peuvent être provisoirement interprétées dans le cadre du codage prédictif (Rao and Ballard, 1999). Dans nos travaux précédents (Alamia and VanRullen, 2019), les chercheurs ont construit un modèle hiérarchique à sept niveaux du cortex visuel avec une connectivité bidirectionnelle mettant en œuvre le codage prédictif. Au sein de la hiérarchie, les niveaux supérieurs prédisaient l'activité des niveaux inférieurs par le biais d'une rétroaction inhibitrice, et les niveaux inférieurs envoyaient l'erreur de prédiction via une excitation feedforward aux couches supérieures pour corriger leur prédiction. Avec des paramètres biologiquement plausibles (constantes de temps neuronales, délais de communication), ce modèle a produit des rythmes voyageant à travers la hiérarchie. Les ondes pouvaient se déplacer dans la direction FW lorsque le modèle recevait des entrées visuelles, et dans la direction BW en l'absence d'entrées (lorsque le modèle traitait des "prieurs descendants" au lieu de signaux sensoriels ascendants).

Dans ce contexte, il est raisonnable de déduire que les ondes FW transportent les signaux d'"erreur résiduelle" (la différence entre les entrées visuelles réelles et la prédiction des régions de niveau supérieur), tandis que les ondes BW transportent les signaux de prédiction. Il est remarquable que les résultats actuels, selon lesquels les ondes FW n'apparaissent que pendant la stimulation visuelle et les ondes BW sont dominantes à l'état de repos, soient en accord avec ce cadre. D'autre part, la corrélation négative dans le temps entre les ondes FW et les ondes BW pendant la stimulation visuelle peut refléter la dynamique du mécanisme

de codage prédictif. En d'autres termes, des signaux de prédiction plus forts dans les ondes BW sont associés à des erreurs de prédiction plus faibles portées par les ondes FW et vice versa. De plus, dans la condition statique, les ondes BW ont augmenté mais les ondes FW ont diminué de manière significative aux derniers stades de la stimulation visuelle, ce qui indique que les informations de prédiction deviennent plus fortes tandis que les signaux d'erreur s'affaiblissent avec le temps. Ce n'était pas le cas dans la condition dynamique, qui présente une structure temporelle de stimulus beaucoup plus complexe (et imprévisible), entraînant des signaux de prédiction moins précis et des signaux d'erreur plus importants.

En résumé, nos résultats suggèrent que les ondes progressives n'apparaissent qu'en présence de stimuli visuels dirigés de bas en haut et disparaissent en l'absence d'entrées visuelles ; les ondes descendantes, bien qu'elles reçoivent une certaine modulation de l'entrée visuelle externe, sont moins affectées. Conformément à l'étude de modélisation de Alamia and VanRullen (2019), les ondes progressives oscillatoires pourraient être une signature neuronale du codage prédictif, les ondes avant transportant les erreurs de prédiction et les ondes arrière transmettant les signaux de prédiction. Il convient de mentionner que, dans l'observation empirique, les signaux de prédiction et d'erreur peuvent se mélanger. Par conséquent, il pourrait être inapproprié de lier directement les signaux biologiques observés aux signaux de prédiction ou d'erreur suggérés par le codage prédictif.

3. Le réseau neuronal biologiquement plausible : réseau de codage prédictif profond

Dans la deuxième étude, nous évaluons les performances du codage prédictif dans les réseaux neuronaux profonds. L'idée est que si le codage prédictif peut servir de principe de traitement sous-jacent dans le cerveau, le modèle mettant en œuvre le codage prédictif devrait, dans une certaine mesure, se comporter comme les humains. D'une part, le codage prédictif a été modélisé par la méthode traditionnelle (voir les sous-sections 1.3.2.1 et 1.3.3.3). Ces modèles sont concis et efficaces, mais ils ne peuvent pas gérer de grands ensembles de données et des tâches cognitives complexes comme les humains, ce qui limite

la comparaison entre les modèles de codage prédictif et les performances humaines. Par conséquent, il pourrait être nécessaire de se tourner vers les réseaux neuronaux profonds alimentés par l'apprentissage profond, car ils sont capables de simuler un grand nombre de neurones et de paramètres et donc capables de réaliser diverses tâches complexes, comme indiqué à la section 1.4.1.

D'autre part, les réseaux neuronaux profonds sont confrontés à leurs propres problèmes. Par exemple, dans les tâches de vision par ordinateur, différents points de vue d'une même image peuvent donner lieu à des jugements complètement différents par les modèles d'apprentissage profond. Le codage prédictif pourrait-il contribuer à améliorer leurs performances ? Ces dernières années, le processus de calcul du codage prédictif a été mis en œuvre dans des réseaux neuronaux profonds (voir section 1.4.2). Ces résultats montrent que le codage prédictif peut apporter de meilleures performances aux réseaux de neurones profonds. Cependant, comme on peut le constater, ces modèles de codage prédictif peuvent varier en termes de formulation mathématique, de construction de modèles et de régimes d'entraînement. Quelle méthode de modélisation est la plus susceptible d'être adoptée par le cerveau biologique?

Dans la deuxième étude, nous avons utilisé le modèle conçu par Choksi et al. (2021). Ce modèle s'est avéré plus robuste vis-à-vis des entrées visuelles bruyantes. Ce modèle reflète-t-il mieux la façon dont le cerveau fonctionne ? Pour répondre à cette question, une façon possible est de tester si un tel modèle peut montrer une perception semblable à celle des humains. Nous testons ici la perception illusoire des contours. Lorsque l'on présente aux humains des images illusoires, comme un carré de Kanisza, en plus des inducteurs de composantes, un carré illusoire peut également être perçu. Les réseaux neuronaux profonds à haute performance ont tendance à rapporter ce qu'ils "voient", c'est-à-dire uniquement les inducteurs. Nous testons ici si l'introduction d'une stratégie de codage prédictif dans les réseaux de neurones profonds peut aider à gagner la perception illusoire.

L'objectif de cette étude était de tester si un réseau neuronal à réaction avec une dynamique récurrente inspirée du cerveau percevrait les contours illusoires (carrés de Kanizsa) d'une

manière similaire à celle des humains. L'augmentation d'un CNN feedforward avec une dynamique récurrente de codage prédictif nous a permis (i) d'analyser les décisions de classification explicites (carré vs inducteurs) et, contrairement à d'autres travaux connexes (Baker et al., 2018 ; Kim et al., 2021 ; Lotter et al., 2018), (ii) de visualiser les entrées reconstruites du point de vue du modèle. Comme indiqué dans une version préliminaire de cette étude récemment publiée dans un atelier de conférence (Pang et al., 2021), nous avons constaté que, par rapport à une ligne de base feedforward, la dynamique récurrente a conduit le réseau à percevoir davantage de contours illusoires. Notamment, en inspectant les reconstructions du réseau, nous avons pu visualiser directement la représentation interne du stimulus par le réseau, ce qui fournit une mesure beaucoup plus claire de la 'perception illusoire' que les travaux précédents. Nous avons trouvé des preuves de modulations des profils de luminance perçus dans la direction attendue pour les formes illusoires, ce qui suggère que le réseau " perçoit " réellement les contours. Nous avons étendu cette analyse et effectué des études d'ablation systématique, tant au moment du test qu'au moment de l'entraînement, et nous avons constaté que le terme de correction d'erreur de rétroaction est essentiel à la perception de l'illusion, tandis que le terme de correction d'erreur d'anticipation tend à la diminuer. De même, l'exploration des ensembles de données utilisés pour le pré-entraînement et le réglage fin du réseau a révélé qu'une exposition préalable aux statistiques des scènes naturelles est un élément crucial de la perception des contours illusoires. Enfin, nous avons également implémenté la dynamique de codage prédictif dans un modèle VGG standard, et nous avons constaté que le modèle PVGG modifié présentait également la perception de contours illusoires. Cela suggère que la perception des contours illusoires découle de la dynamique de rétroaction du codage prédictif, indépendamment de l'échelle du modèle. En résumé, nous fournissons des preuves claires que la dynamique récurrente inspirée par le cerveau peut amener les réseaux à percevoir les contours illusoires comme les humains.

Bien qu'il existe des différences intrinsèques entre le système visuel humain et les réseaux neuronaux artificiels (par exemple, les signaux d'erreur globaux nécessaires à l'apprentissage par rétropropagation (Lillicrap et al., 2020), nous soutenons que les résultats actuels mettent

en évidence trois similitudes essentielles avec la vision biologique. Premièrement, les deux systèmes peuvent effectuer un traitement global similaire des contours illusoires. Dans le système visuel, Pan et al. (2012) ont signalé que les contours illusoires activent des représentations équivalentes dans V4 par rapport aux contours réels, alors que V1 et V2 encodent différemment leurs caractéristiques locales respectives. Autrement dit, en plus du traitement local dans les premières régions visuelles, il existe un mode de traitement global par lequel les inducteurs illusoires forment des représentations intégrales des contours. De façon similaire, lorsqu'on lui présente des contours illusoires, le réseau actuel attribue une probabilité beaucoup plus élevée à la classe " carré " que pour les images de contrôle aléatoires ou tout venant, bien qu'elles partagent les mêmes caractéristiques locales que les images de contours illusoires tout venant. Cela indique que le réseau possède également une capacité de traitement global. De plus, ce traitement global résulte principalement des connexions de rétroaction, car aucun des réseaux de rétroaction testés n'a pu percevoir les contours illusoires (tableau 3). Deuxièmement, la performance "comportementale" (c'est-à-dire la probabilité de décision) du réseau est également cohérente avec la recherche physiologique sur les contours illusoires. Lee et Nguyen (2001) ont comparé l'activité EEG pour les contours illusoires et d'autres motifs, et ont constaté que l'activité pour les contours illusoires est significativement plus élevée que les stimuli aléatoires de contrôle, mais toujours inférieure aux contours réels. Dans la présente étude, les probabilités de classe "carrées" attribuées par le réseau après la couche Softmax indiquent un schéma similaire. Enfin, au niveau " perceptuel ", nous avons vérifié directement la représentation interne de la première couche du réseau (par sa voie générative de " reconstruction d'image "). La métrique FG suggère que le réseau perçoit une forme illusoire plus claire (ou plus sombre), ce qui est cohérent avec la " luminosité illusoire " rapportée lorsque les humains perçoivent des contours illusoires (Parks, 2001 ; Schumann, 1918 ; Spillmann Dresp, 1995).

L'étude suggère que le réseau neuronal piloté par la dynamique du codage prédictif possède une perception illusoire, soutenant la possibilité que la même stratégie dynamique, c'est-à-dire le codage prédictif, puisse être partagée entre le réseau et le cerveau biologique.

4. Ondes progressives dans le réseau de codage prédictif profond

Les deux premières études nous ont appris que: (i) les ondes progressives corticales peuvent refléter les mécanismes neuronaux du codage prédictif dans le cerveau biologique, les ondes vers l'avant véhiculant les erreurs de prédiction et les ondes vers l'arrière transmettant les prédictions ; (ii) le réseau neuronal profond qui met en œuvre le codage prédictif pourrait percevoir des contours illusoires comme les humains, ce qui signifie que ce schéma peut sous-tendre certaines fonctions cérébrales, sinon toutes, comme la perception illusoire. Une question naturelle peut se poser : un tel réseau neuronal, partageant la même architecture fonctionnelle et la même dynamique de signal que le cerveau humain, pourrait-il reproduire ses oscillations ou même ses ondes progressives ? Si oui, nous pourrions obtenir des preuves à l'appui de la première étude suggérant que les oscillations ou les ondes progressives sont effectivement les corrélats neuronaux du codage prédictif, ainsi que de la deuxième étude soutenant le rôle du codage prédictif comme modèle du cerveau humain.

Dans la troisième étude, nous mettons à jour le même réseau neuronal prédictif que dans la deuxième étude en ajoutant des délais et des constantes biologiquement plausibles entre les couches du réseau afin de générer des oscillations. Nous évaluerons d'abord si un tel modèle peut générer des oscillations et des ondes progressives biologiquement plausibles entre les couches. Une fois que nous aurons obtenu les résultats souhaités, dans un deuxième temps, nous pourrons utiliser ce réseau neuronal oscillatoire comme modèle de travail du cerveau physique et nous pourrons l'utiliser pour tester un phénomène physiologique ou psychologique peu clair ou controversé en vérifiant soigneusement les activités des couches ou des neurones artificiels, ce que nous ne pouvons pas facilement faire dans le cerveau biologique.

Les résultats préliminaires de la troisième étude montrent que le réseau pourrait osciller avec des paramètres temporels biologiquement plausibles. Nous nous attendons à ce qu'un tel réseau neuronal oscillatoire produise des résultats plus proches de ceux de l'homme en termes de modèle d'activation des unités de signal et de sortie de décision finale.

4. Ondes progressives dans le réseau de codage prédictif profond

La présente thèse a tenté de résoudre ces questions en évaluant la théorie du codage prédictif à la fois dans le cerveau biologique et dans les réseaux de neurones profonds. D'une part, nous avons essayé de trouver la dynamique neuronale dans le cerveau qui se rapporte aux signaux de prédiction descendants d'un niveau supérieur vers un niveau inférieur ainsi qu'aux signaux d'erreurs de prédiction ascendants qui se manifestent lorsqu'il y a inadéquation entre la prédiction et les preuves observées. D'autre part, nous avons tiré parti des réseaux neuronaux profonds qui sont censés imiter vaguement la structure du cerveau et les comportements humains tels que la vision par ordinateur. L'idée est que si le cerveau adopte une stratégie de codage prédictif pour alimenter ses fonctions, un réseau neuronal profond qui met en œuvre une dynamique de codage prédictif devrait également présenter des performances similaires à celles de l'homme.

Plus précisément, nous avons mené trois études progressives et connexes. La première se concentre sur les activités neuronales dans le cerveau biologique, en particulier sur les ondes progressives oscillatoires. Les ondes progressives corticales pourraient-elles sous-tendre les processus prédictifs dans le cerveau ? Nos résultats suggèrent que les ondes progressives n'apparaissent qu'en présence de stimuli visuels ascendants et disparaissent en l'absence d'entrées visuelles ; les ondes descendantes, bien qu'elles soient modulées par des entrées visuelles externes, sont moins affectées. Nous avons expliqué les résultats dans le cadre du codage prédictif : les ondes progressives oscillatoires pourraient être une signature neuronale du codage prédictif, les ondes avant transportant les erreurs de prédiction et les ondes arrière transmettant les signaux de prédiction.

Dans le second projet, nous avons utilisé la technique de l'apprentissage profond et construit un réseau neuronal capable de mettre en œuvre la dynamique du codage prédictif. Si le cerveau humain fonctionne selon un tel principe dynamique, le réseau de codage prédictif devrait présenter certaines propriétés semblables à celles de l'homme. Nous avons testé cette hypothèse avec des stimuli d'images de contours illusoires. L'étude suggère que le

réseau neuronal piloté par la dynamique du codage prédictif possède une perception illusoire, soutenant la possibilité que la même stratégie dynamique, c'est-à-dire le codage prédictif, puisse être partagée entre le réseau et le cerveau biologique.

Sur la base des résultats des deux premières études : (i) les ondes progressives corticales reflétant les mécanismes neuronaux du codage prédictif dans le cerveau biologique ; (ii) le réseau neuronal prédictif affichant des performances semblables à celles de l'homme, dans la troisième étude, nous mettons à jour le même réseau neuronal prédictif en ajoutant des délais et des constantes de temps biologiquement plausibles entre les couches du réseau afin de générer des oscillations. Les résultats préliminaires montrent que le réseau peut osciller avec des paramètres temporels biologiquement plausibles. Nous pensons qu'un tel réseau neuronal oscillatoire produira des résultats plus proches de ceux de l'homme en termes de modèle d'activation des unités de signal et de sortie de décision finale.

En résumé, la thèse affirme la possibilité de la théorie du codage prédictif comme cadre unificateur des fonctions cérébrales en combinant les preuves du cerveau biologique et du réseau neuronal computationnel.

1 Literature Review

1.1 General Introduction: An overview

When presented with a cat picture, how does our brain perceive it as a cat representation? Intuitively, the cat perception can be formed simply by a forward message flow as shown in Figure 1A. However, an influential theory in the field of neuroscience termed 'predictive coding' holds an opposite opinion: instead of passively receiving and being driven by external stimuli, our brain holds an inner model predicting the outside world continuously as shown in Figure 1B. The predictive coding theory not only can explain the formation of visual perception, instead, it can also serve as a unifying framework for brain functions like memory, motor control, reasoning, etc. In this chapter, I will give a comprehensive introduction to this theory, especially under the context of physical brains and deep neural networks.

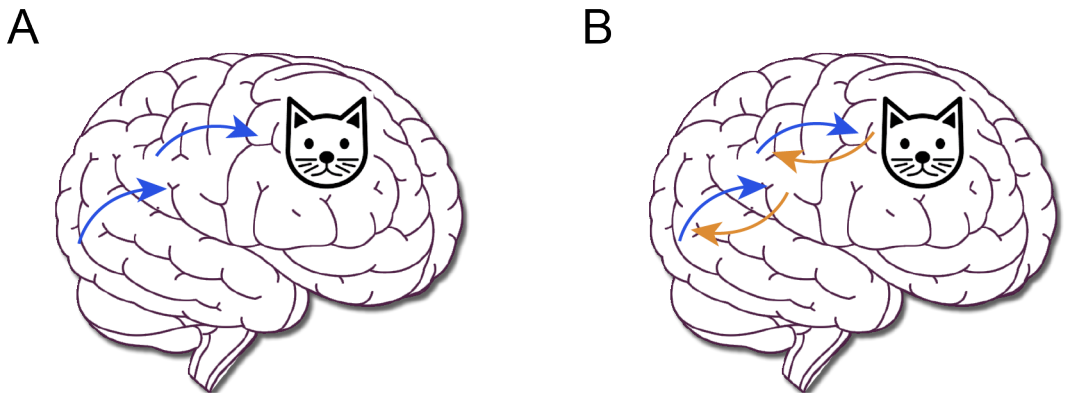


Figure 1: **Visual perception.**

Before introducing the predictive coding theory of brain function, I will first describe in section 1.2 the overall structure of the human brain for a better understanding of how a particular processing strategy postulated by a theory could perform within the brain structure. Anatomically, the brain cortex is organized in a hierarchy where forward connections lead from lower regions to higher regions; while backward connections are projected in an opposite direction. Functionally, the processing in the hierarchy is considered topological, modular,

and progressively complex. Given such a structure, different theories appeared attempting to describe the brain function in a single and simple framework, such as the Global Workspace Theory (GWT), Bayesian brain framework, etc.

Bayesian brain framework only provides a theoretical description of the processing in brains, while predictive coding theory gives the concrete computations reflecting a Bayesian brain. In section 1.3, I will state the core idea of predictive coding theory as well as how the theory explains a range of physiological and psychological facts about the biological brain. Importantly, as a guide of brain computation, predictive coding has been formalized by several researchers. Thus I will show how those formalized algorithms could model the neurophysiological facts in the brain.

The deep learning technique provides a new avenue for the implementation of predictive coding algorithms since it allows the algorithm to learn in a more human-like way. In section 1.4, I will give a brief introduction on how deep neural networks could be implemented in terms of model construction, training, and testing. Especially, I will highlight one type of deep network model called Convolutional Neural Network (CNN) which is the one we adopted in the current work. Besides, I will compare several deep neural networks which implement predictive coding dynamics in terms of their construction, learning methods and testing tasks.

In section 1.5, I will discuss the neural mechanisms of predictive coding in biological brains. Especially, whether cortical oscillations could serve as a possible candidate. To figure it out, I will first present the basic properties of oscillations as well as their corresponding cognitive functions. Interestingly, both predictive coding and cortical oscillations are thought to involve a wide range of neurophysiological phenomena. Then I will point out a special neural event, the traveling waves, which belong to cortical oscillations by nature but can propagate across brain regions. It seems that the propagation property of traveling waves can be a signature of predictive coding since the traveling oscillations could convey necessary signals by predictive coding.

Lastly, I will summarize this Chapter in section 1.6 in terms of two aspects as suggested in the thesis title, i.e., in the brain and also in the deep neural networks. Based on that, I will raise three specific questions that the current thesis needs to address.

1.2 The hierarchy in brain structure and function

As we know, structure determines function. Therefore, to understand a complex system, one first needs to figure out its arrangement or constitution and based on it to further explore how each component behaves to generate collective activity. The same is true for biological brains. What are the inner workings of the brain that enable it to have various functions, such as sensation, perception, attention, memory, etc? In this section, I will give a brief introduction to the brain structure showing how its inner components are organized and connected. Before that, let me first show what we have in the brain.

At the macro level, the brain could be divided into three parts: brainstem, cerebellum, and forebrain. As the largest region, the forebrain could be divided into two cerebral hemispheres, connected by the corpus callosum. The outer layer of the forebrain is called the cerebral cortex which plays an important role in brain functions. In addition, the thalamus is also an important part of the forebrain. Most of the received sensory information will be relayed by the thalamus to various cortical regions for further processing and integration, except olfactory signals.

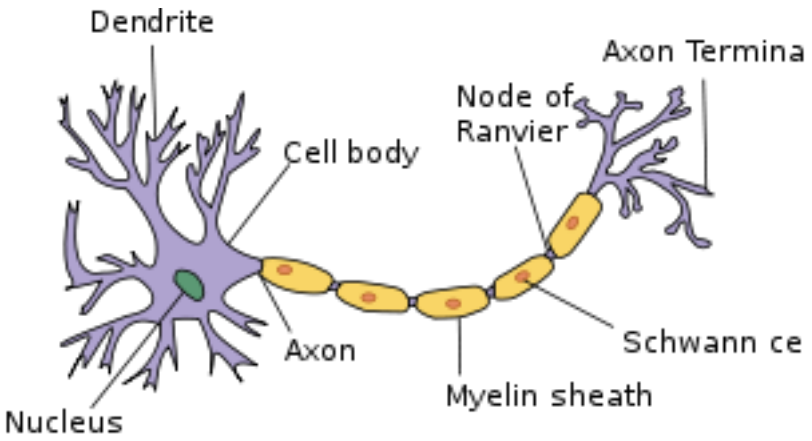


Figure 2: **A prototypical Neuron.** It basically contains three parts: cell body, dendrites and axon. Figure from: https://upload.wikimedia.org/wikipedia/commons/b/b5/Neuron.svg.

From a microcosmic point of view, the most basic structural and functional units in the

nervous system are neuron cells. The cerebral cortex in humans contains approximately 14 to 16 billion neurons. Figure 2 shows a prototypical cell, which consists of a cell body, dendrites, and a single axon. The cell body could integrate input information received by dendrites which are short and include many branches. The integrated information is then sent out by the long axon to another neuron. An important structure for information transmission between two neurons is called a synapse. The synapse can be excitatory or inhibitory. An excitatory synapse enables the presynaptic neuron to increase the activity in the postsynaptic neuron.

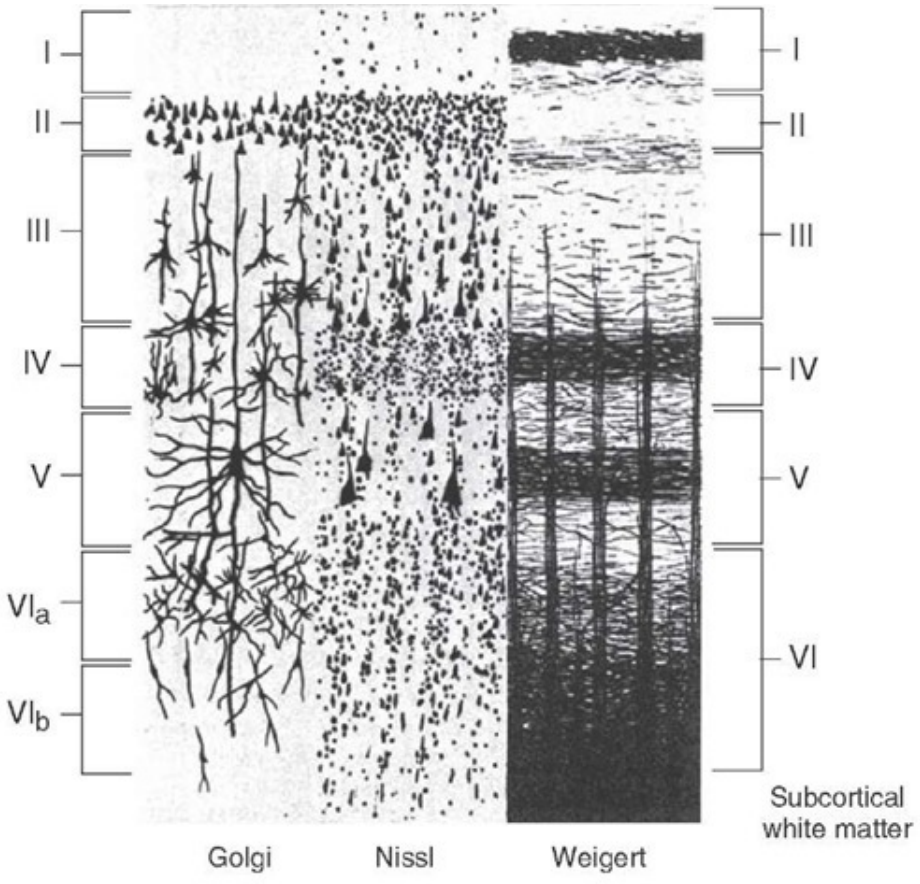


Figure 3: **Six layers of neocotex.** Figure from:https://epomedicine.com/wp-content/uploads/2016/07/cortex-layers.jpg.

One characteristic of the cerebral cortex is its laminar structure. The neocortex, occupying

around 90% of the cerebral cortex is formed of six distinct layers, as shown in Figure 3. These layers are numbered I to VI from the outermost to the innermost layer and they differ in the cell's density, size, shape, etc. There is another way to sort those layers. Layer IV can be termed as granular layer; the layers above can be grouped together to be referred to as supergranular layers; the layers below are thus the infragranular layers. Except for neuronal content, cortical laminae can also differ in their connectivity to other parts of the brain. This will be discussed in the following section.

Another characteristic is the columnar organization of cortices. The cortical column is also called the functional column which describes the functional connection of the six-layer cortex in the vertical direction. A column of neurons will tune to the same set of features. Therefore it has been proposed that neurons within the column may constitute a fundamental computational/functional unit of the cerebral cortex (Mountcastle, 1997).

1.2.1 The hierarchical connectivity of brain cortex

Now we have a preliminary understanding of the human brain with the above-mentioned constituent components and some basic characteristics. A natural question arises: How does the brain arrange those parts into a whole, complex system which can perform sensory processing, cognition, and control. In practice, it is not realistic to check the anatomical configuration of the whole brain, ranging from inter-neuronal to inter-regional connectivity, due to its huge number of neurons stored in the brain (not to mention the number of synapses). Scientists usually examine a smaller but representative region to find its pattern and then generalize it to other regions or even the whole brain.

In history, the cortical sensory areas, especially the visual area are the most studied. Human beings use five senses to collect outside information. Among them, vision is the most important one, since the visual system extends over a relatively larger area than other sensory systems. It is therefore a particularly relevant system to study in order to understand the

whole brain.

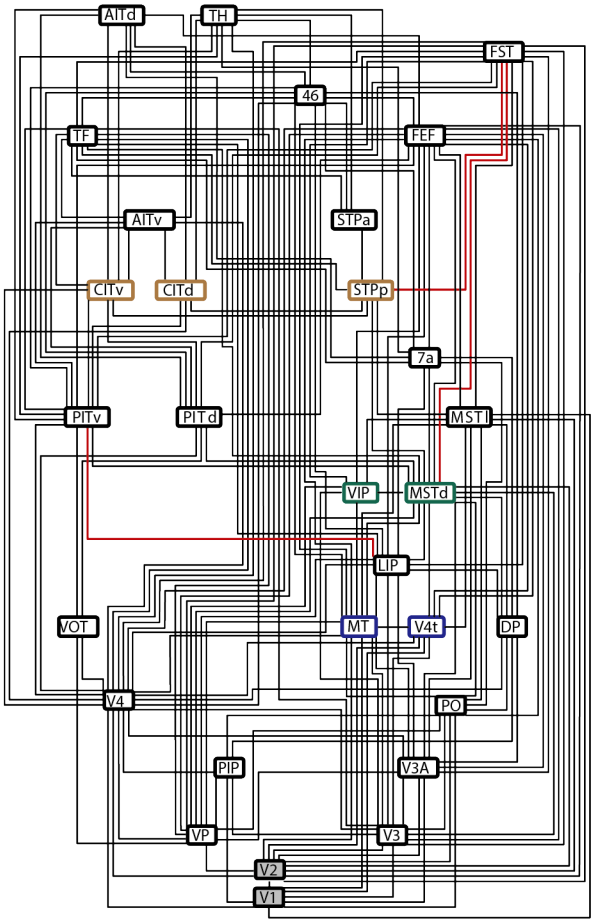


Figure 4: **The wiring diagram in visual cortex.** The brain areas are arranged by their laminar origin and termination patterns of projections between each other. Figure from Hilgetag and Goulas (2020)

In the visual system, there are very rich connections between different cortical areas. Early anatomical studies examined the laminar patterns of the cortico-cortical projection origins and termination, according to which, three projection directions could be defined (Rockland and Pandya, 1979; Felleman and Van Essen, 1991): (i) forward projections, meaning the projections originate most from supragranular cortical layers and terminate in the granular layer (layer IV); (ii) feedback projections, originating predominantly from deep infragranular cortical layers and targeting superficial and deep layers; (iii) lateral projections, originating

from upper and deep cortical layers in a more equal proportion and terminating in the target areas in a column-like fashion across all layers.

The above-mentioned three types of cortico-cortical connections could be viewed as extrinsic connections in terms of the concept of the functional column. Inside the column, the inter-neuronal connectivity could be viewed as intrinsic connections. It bears mentioning that the extrinsic connections only make up a very small amount of connections in the cortex; the majority of connections come from the local intrinsic connection (Douglas and Martin, 2012). It has been proposed that such intrinsic connections could be cast as a canonical microcircuit (Mountcastle, 1997).

The extrinsic cortico-cortical connections in the visual system form the serial or hierarchical processing (Van Essen et al., 1992). The notion of hierarchical configuration was then extended to other systems including somatosensory (Felleman and Van Essen, 1991; Iwamura, 1998) and auditory (Kaas and Hackett, 1998) areas. The hierarchical configuration in the visual system can be generalized to the whole brain system. Meunier et al. (2009) performed a hierarchical modular decomposition of large numbers of brain functional networks. Their study revealed the largest five modules at the highest level of the hierarchy: medial occipital, lateral occipital, central, parieto-frontal, and fronto-temporal systems. These roughly match the hypothesized hierarchies proposed by Mesulam (1998) (See Figure 5).

1.2.2 The neural processing in the Hierarchy

In the last section, I explained that the brain might be organized in a hierarchical approach. In this part, I will try to show some principles that such a hierarchical system might use. Since the visual system is studied extensively, I will illustrate that mainly in the context of the visual system.

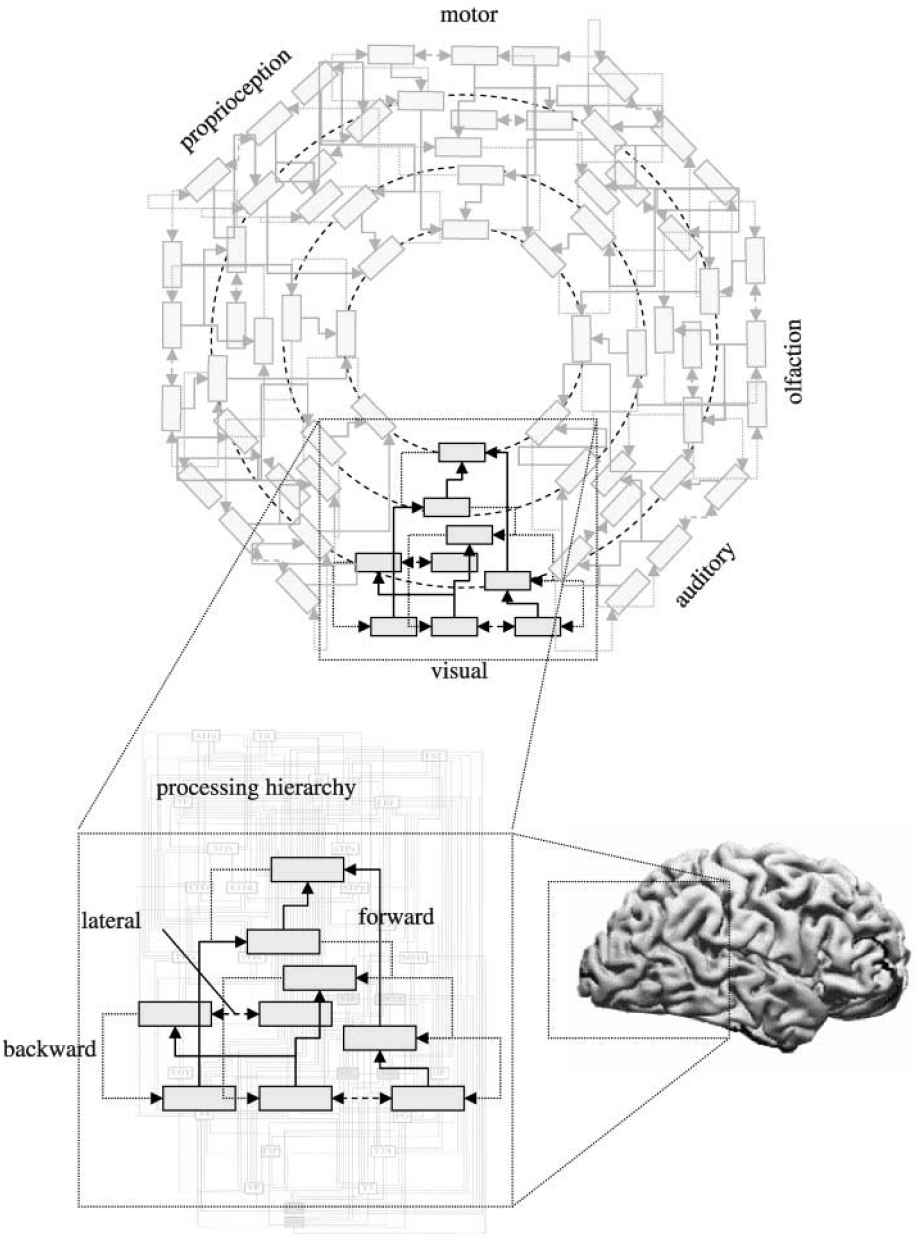


Figure 5: **The sensory brain is organized in hierarchies.** Figure from (Mesulam, 1998; Friston, 2005)

1.2.2.1 Retinotopy or topology in the hierarchy

Retinotopy is a processing property of cells in the hierarchy of the visual system starting from the retina ascending upwards through the entire visual cortex. It refers to the topological way of mapping visual input in the visual field to each level of the visual stream. That is,

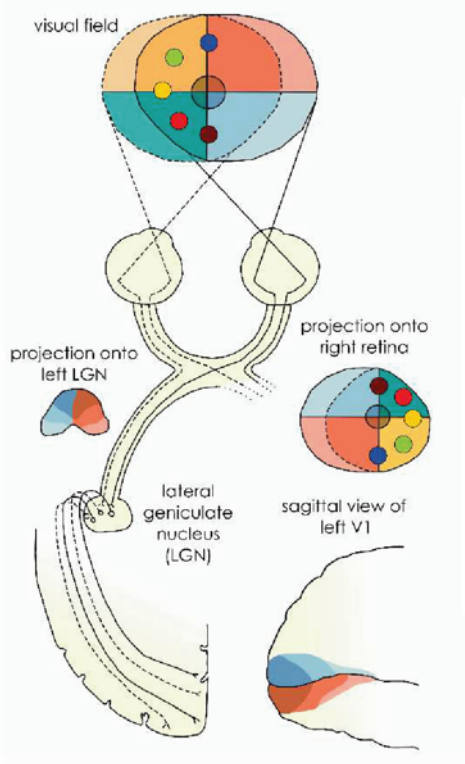


Figure 6: **Retinotopic mapping in the visual system.** The retina receives a projection of the visual field in the outside environment onto its photoreceptors. Two spatially close points in the visual field will activate two close-by photoreceptors in the retina. Through the optic nerve, this projection continues via LGN to V1 while keeping the same spatial relationship between neurons. Figure reproduced from Wallisch and Movshon (2008).

neighboring information in the visual field, for example, two patches in a picture, is processed by spatially close cells in the brain. For clarity, the retinotopic organization of the visual system could be viewed as a special case of a more general topographic organization. Thus the words 'retinotopic' and 'topographic' sometimes are used exchangeably in the literature.

Many brain structures along the visual stream are organized into retinotopic maps. However, as ascending the visual hierarchy, the retinal mapping becomes complex. The retinotopic maps of brain areas up to V1 along the visual stream can be cast as first-order representation, which means the mapping is straightforward and continuous. For example, nearly located cells in the retina will project to neighboring cells in V1 as shown in Figure 6. As one ascends the visual stream beyond V1, the retinotopic maps become more and more complicated

and therefore are called second-order representation. Therefore, the retinotopic organization becomes less pronounced as ascending to higher visual areas (Hilgetag and Goulas, 2020).

1.2.2.2 Across the hierarchy: parallel/distributed/modular representations

Simon (1962) suggested that most complex systems are organized in hierarchical modularization or modular-in-modular. We already know that there are several top-level modules or systems in the brain: the visual system, auditory system, olfactory system, etc. According to the inference of Simon (1962), to form a complex brain system (we already know it is complex), there should exsit submodules in the top-level modules and subsubmodules in the submodules and so on. Since a module has the meaning of encapsulating information, each module can process information in a relatively independent way and thus achieving parallel or distributed processing which is more efficient than serial processing.

In the visual system, it has been well established that there are two visual streams (or two modules) in visual processing: 'what' and 'where' streams. As the name suggests, 'what' streams process recognition-related information, while 'where' streams process location-related information. The two streams interconnect at V1 and then separate afterwards into two separate pathways: The ventral pathway for the 'what' stream and the dorsal pathway for the 'where' pathway (See Figure 7).

The evidence for the two-streams hypothesis mainly came from lesion studies or cortical impairments studies. Mishkin and Ungerleider (1982) reported in their seminal study that the bilateral removal of the posterior parietal cortex resulted in severe impairments of landmark discrimination in a task where monkeys had to choose a closer covered foodwell relative to a movable cylinder on the table. This study proved a critical role of the parietal cortex as the 'where' pathway in spatial and reaching abilities. The two-streams hypothesis was first proposed by Mishkin and Ungerleider (1982); Mishkin et al. (1983) originally stating the recognition properties for the 'what' pathway and spatial properties for the 'where' pathway.

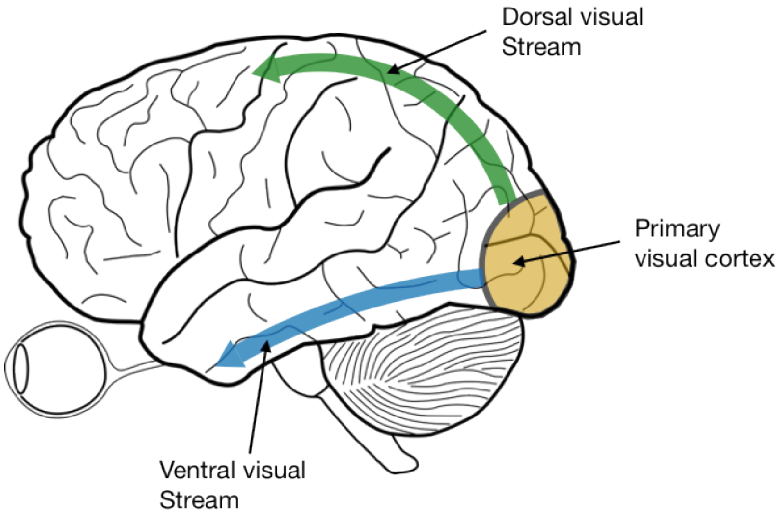


Figure 7: **Parallel processing streams in visual system.** Originating from a common source in the visual cortex, visual information is divided into two streams: the dorsal stream (green) and the ventral strem (blue). Figure from Boutin (2020).

According to the inference by Simon (1962), the above-mentioned parallel processing pattern should be common in the brain's hierarchical organization. This may be true. First, although the two visual streams interconnect at V1, their starting location may be as early as the retina. Livingstone et al. (1988) investigated the response properties of ganglion cells in the retina and reported that based on the specific types of ganglion cells, two main sets of visual information could be shifted to the visual cortex in a parallel way: the first set of information is about depth, motion and the other one is color and object recognition. Second, the two main streams should contain the substreams that could deal with even more detailed and specific information. Third, such distributed organization also exists in other modules or system. It has been reported that there also exists two processing streams in the auditory system (Eyesenck and Keane, 2010). Therefore, the modularity of the hierarchical structure may be a common way of organization.

1.2.2.3 Along the hierarchy: Progressive feature detectors

In the last section, I have made a horizontal comparison between and inside the hierarchies and showed the modular nature of the hierarchical structure in the brain. In this section,

I will evaluate from a vertical perspective and check whether the anatomical hierarchy is related to hierarchical processing as put forward by Hubel and Wiesel (1962). I will focus on the ventral pathway or the 'what' stream and examine the processing properties of each level in ascending order in its hierarchy.

Anatomically, along the ventral pathway in ascending order, we have the retina, Lateral Geniculate Nucleus (LGN), primary visual cortex (V1), the secondary visual cortex (V2), the quaternary visual cortex (V4), the inferior temporal cortex (IT). One metric for estimating the processing characteristic is Receptive Field (RF). In the visual domain, the RF can refer to the limited spatial area in the visual field. The size and complexity of RF for neurons at different levels can be different.

From retina to LGN

The receptive fields of ganglion cells have a circular shape. It has been reported that two types of receptive fields exist: On-center and Off-center receptive fields as shown in Figure 8. The cell that bears an On-center receptive field will increase the firing when the light increases in the central disk with no light in the surrounding concentric ring. For an Off-center cell, the opposite is true. That is, light in the surround will increase the firing of the cell while light in the center might decrease its firing response.

With such receptive fields, the ganglion cells can not only transmit light information but also contrast signals. Researchers have used it for detecting edges (Higgs, 2014). Further along the visual hierarchy, groups of ganglion cells form the receptive fields of LGN cells, whose receptive fields show the same property as ganglion cells.

From LGN to V1

Compared to ganglion and LGN cells that are sensitive to small and simple stimulation of light spots, Hubel and Wiesel (1959) found that V1 cells could respond to relatively larger and complex visual features. Different V1 neurons could be elicited by differently oriented slits of light in their receptive field. Remarkably, by checking the property of the receptive

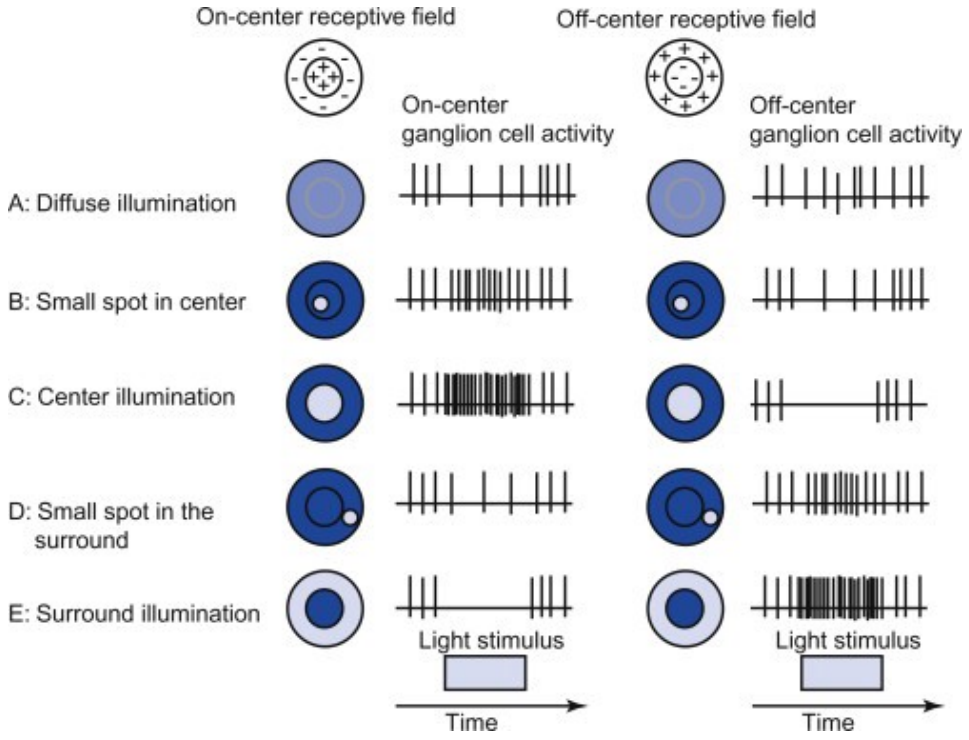


Figure 8: **Receptive fields of retinal ganglion cells.** Ganglion cells with On-center and Off-center receptive fields will respond oppositely to light spots in the center or surround. Figure from Kiley and Usrey (2013)

field, Hubel and Wiesel (1962) build the relationship between LGN neurons and V1 neurons. The integration of LGN neurons' RF could result in the specific preference of oriented light bar for neurons in V1 as shown in Figure 9.

Their further investigation in V1 neurons led to the discovery of three types of cells: simple, complex, and hypercomplex cells. These cells correspond to receptive fields with increasing complexities and size. Simple cells are selective to bars that oriented with specific angle and position which sit in their relatively small receptive fields. Complex cells possess relatively bigger receptive fields. They are still selective to oriented bars but become insensitive to the bars' position, meaning the acquisition of invariance to position. the hypercomplex cells (also called the end-stop cells) only respond to bars with certain length and stay invariant to other properties of the bar, meaning an even more complex receptive field.

Those facts suggest: (i) Each level receive inputs from their immediate predecessor; (ii)

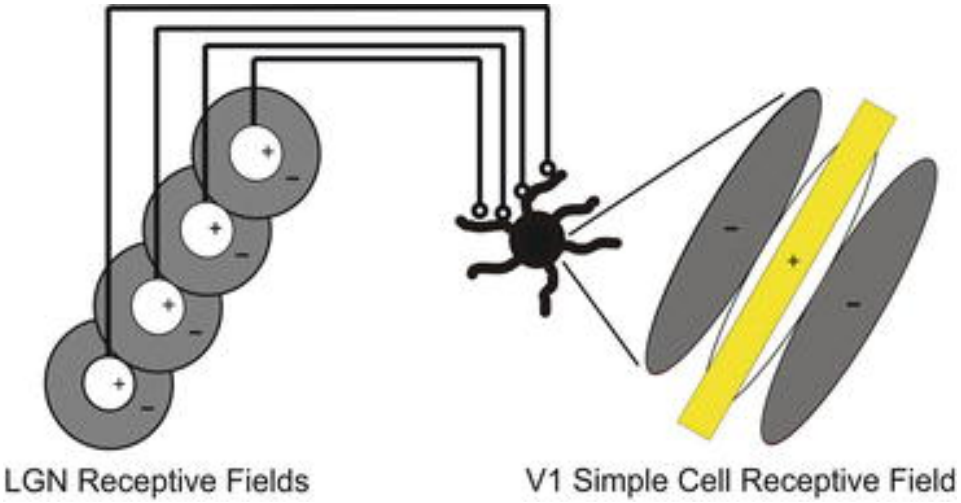


Figure 9: **Receptive fields integration from LGN to V1.** Figure from Kiley and Usrey (2013)

The visual hierarchy encodes information abiding by the principle from simple to complex, indicated by the progressive increase in the complexity of receptive field properties; (iii) This principle also consists of the principle of distributed representation since a cell in higher-level receives information from multiple cells in lower level.

In V4

As one moves to the intermediate areas in the visual stream such as V3 and V4, the receptive fields become even larger and complex. David et al. (2006) reported that V4 neurons were selective for spirals and concentric curves. That is, compared to V1 neurons, which are selective for simple straight lines; V4 neurons are better predicted by non-linear configuration. Moreover, David et al. (2006) provided evidence for the bi-modal tuning for some V4 neurons. In other words, V4 neurons are responsive to a combination of differently oriented bars, meaning the components of shapes. The studies by Pasupathy and Connor (2002) also provided supportive evidence showing that V4 neurons could encode shapes based on their constituent boundary features or contours which are represented in an object-centered manner.

In IT

The seminal work by Gross et al. (1972) and later work Desimone et al. (1984) have found evidence of selective responding to perceptually meaningful complex stimuli, such as faces and other objects in inferior temporal (IT) neurons. Subsequent neuroimaging research identified several category-specific areas in IT: Fusiform Face Area (FFA) for faces (Kanwisher et al., 1997); Visual Word Form Area (VWFA) for visual forms of words (Epstein and Kanwisher, 1998); Extrastriate Body Area (EBA) for body parts (Cohen et al., 2000). For IT neurons, a series of works by Keiji Tanaka have been carried out to investigate the properties of their object selectivity. Visually similar stimuli such as different views of an object could be organized in columns. As one moves along the ventral visual pathway, such viewpoint selectivity of neurons in IT decreases; instead, the tolerance or invariance for an object's viewpoint increases, meaning a stable representation for an object could form with a random viewpoint.

In summary, it seems that the cells' response characteristics (reflected from their receptive fields) change systematically as one moves to higher-level areas within the hierarchical structure. The cell at a given stage of the hierarchy receives information from many cells located one level below and after processing and aggregation. The output from this cell and other cells in the same level forms the input for the next cell located one level higher. Consequently, the receptive fields become larger and the processed features transform from simple, local, and concrete to complex, global and abstract.

1.2.3 Theories of brain

Now we have learned the basic knowledge on brain structure. The brain regions are connected with each other in a hierarchical fashion. Moreover, the inside neuronal representation also present hierarchical properties with increasing complexity and abstractness along with the hierarchical structure. How are various brain functions realized based on such a structure? Many theories have been proposed to explain how information flow in the brain is routed

and processed to generate those functions.

For instance, the Bayesian brain framework attempts to describe how human brains function and interact with the external physical world. The main issue it discusses is around the concept of 'uncertainty'. The physical world is riddled with uncertainties, such as the varying weather conditions or the forthcoming stimulation. How does the brain handle these harsh situations and survive and improve the possibility of survival? The Bayesian brain hypothesis states that our brain holds a hidden internal model of the external world and tries to predict what happens next. Moreover, the Bayesian statistics, especially the Bayes' Theorem (as shown in Equation (1)) provide a simple approach linking the known evidence and the future states (Clark, 2013). That is, our brain can predict the future state based on an internal probabilistic model which can be updated by the outside world.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

In Equation (1), B could represent the outside observation or external world, and A represents our inner parameters or the inner model of the world. Therefore $P(A|B)$ is posterior, $P(B|A)$ is the likelihood, $P(A)$ is the prior and $P(B)$ is the evidence. Our brain tries to increase the possibility of posterior. In the next section, I will introduce the predictive coding theory, which could be viewed one possible way to realize the Bayesian Brain.

1.3 A Unifying Brain Theory: Predictive Coding

In the last section, I mentioned the Bayesian brain hypothesis, which is a theoretical framework to guide the brain's operation. In this section, I will introduce a concrete computational algorithm, termed predictive coding as formulated by Rao and Ballard (1999) which can realize the Bayesian brain operation.

1.3.1 What is predictive coding?

When presented with a cat picture, how does the brain form the representation of a cat instead of a dog? Intuitively, one might suppose the sensory brain consists of feature detectors at each level of the hierarchy, first looking for edges and bars, then shapes until forming a conceptual representation. However, instead of being a passive analyzer, the brain actively infers the most likely causes of the perceived stimuli. The sensory processing by the brain is actually an inference problem. Consequently, the brain needs to solve an inverse problem via a hidden inner model. As pointed out by Spratling (2017), the inverse problems are generally ill-posed, meaning multiple solutions or causes can exist for a perceived stimulus. For better inference, the inner model needs to be constrained and predictive coding offers a strategy for this problem.

In a hierarchical inner model implementing a predictive coding strategy, each level attempts to predict the lower-level representation or the external stimulation for the lowest level through feedback connection. This prediction will be compared with the received ground truth and therefore mismatches or the prediction errors will be computed at the lower level. Through feedforward connection in this model, prediction errors are sent upward to correct the activity at a higher level for better predictions next time. The theory emphasizes the interaction between the top-down predictions and the bottom-up drives, not just the sensory reality from the outside world. Incredibly, the theory points out that the brain function could be achieved merely based on a simple computational process: the minimization of prediction

error.

1.3.2 The evidence for predictive coding in the brain

Although predictive coding has a very concise logic, it is believed to explain a wide range of biological phenomena. Therefore, it might serve as an all-encompassing theory about what the entire brain is doing.

1.3.2.1 Accounts for neuronal characteristics in visual system

Predictive coding has a long record in successfully modeling the responsive pattern of low-level neurons, especially in the visual system.

The center-surround inhibition refers to a responsive pattern in bipolar and ganglion retinal cells. According to this, retinal cells could be divided into On-center and Off-center cells. It was observed by Kuffler (1953): On-center cells respond strongly (meaning fire rapidly) to the center-lighted receptive field; while Off-center cells are activated most with the surround-lighted receptive field. Srinivasan et al. (1982) used predictive coding to model this center-surround inhibition and found it could provide a good interpretation for this phenomenon: retinal cells could represent the prediction error by suppressing the predicted light amount from the perceived one. Thus, Predictive coding could provide a mathematical explanation of the observed spatial decorrelation in retinal cells activation.

The temporal decorrelation observed in the LGN neurons can also be explained by a variant theory of predictive coding (Dan et al., 1996). Visual signals in the natural environment can be highly redundant in spatial and temporal domains. For efficient coding, it has been proposed that the retina and the LGN can remove the redundancy. As stated above, retinal cells can remove spatial redundancy. For LGN neurons, instead of representing the content of visual input, they can only deliver the different information in terms of time to the visual cortex, which can be considered as the prediction errors signals.

The seminal work by Rao and Ballard (1999) used predictive coding to model another extra-classical Receptive Field effect, the end-stopping mechanism. It describes the phenomenon that a cell that responds optimally to an oriented bar will reduce its activity or firing rate when the observed bar extends beyond the cell's classical receptive field. That is, the oriented bar occupied both the extra classical and classical receptive fields. They trained the predictive coding with natural images, which means the model learned knowledge about the statistical regularities in natural images as humans do. During testing, a short line (shorter than the line that existed in the natural images) was presented to the model, it showed a big prediction error, consistent with the brain neuron's activity. Just like the biological neurons, when a deviation from natural statistics is detected, the model could show an end-stopping selectivity. Thus, such selectivity is thought to encode for prediction error.

1.3.2.2 Accounts for evoked cortical responses

The temporal dynamics of a predictive coding system can be described by the constant exchange of predictions and error signals. As a theory of brain function, predictive coding dynamics were related to the temporal course of various evoked cortical responses by Friston (2005). The evoked cortical responses can include the activation pattern of signal neurons, the immediate changes of EEG recordings, or even much slower changes in BOLD signals of fMRI studies. Notably, he pointed out that the evoked responses should not be viewed as either predictions or error signals, since all signals are supposed to mix together in the brain. Therefore the attenuation of cortical responses can be viewed as an observable expression of a process where the brain attempts to minimize the prediction errors generated from the comparison between the prediction of the stimulation and the received stimulation. The decrease of cortical response thus means the representation for the most likely cause of the evoking stimulus (Friston, 2005).

Friston derived two perceptual processes from the single computational dynamics in predictive coding. When the inner model tries to predict the cause of stimulation without changing its connecting weights (or the synapse efficacy in the biological brain), it's the process of

perceptual inference; when the same predicting process is accompanied by the updates of the inner model (or the changes of synapse efficacy), it's the process of perceptual learning. An example of perceptual inference employed by Friston comes from the study on the global precedence effect which refers to a faster response towards the global feature relative to the local one of a given stimulus. In this study, the authors reported a greater posterior N2 component in the evoked ERPs in the incongruent condition relative to the congruent one (Han and He, 2003). Friston explains it within the predictive coding framework: Top-down predictions offer global contextual information in terms of the local low-level activation. When the incongruency occurs between them, evoked cortical response could be observed. Therefore, the evoked response could be understood as a failure suppression of predictions to the prediction errors (Friston, 2005).

Friston's views are supported by subsequent research. For instance, N170, an ERP component, may serve as a brain activity index for the probing of prediction error signals in the tasks like Johnston et al. (2017); Robinson et al. (2020). N170, as an ERPs component, is a negative inflection along with the temporal profile of ERPs activity. It usually occurs ∼150-200 ms after the stimulus onset and has been proven to be a robust and highly replicable index of the visual processes for specific visual objects, such as face (Eimer, 2011; Liu et al., 2000; Rossion and Jacques, 2008), visual word-forms (McCandliss et al., 2003) and danger signals (Levita et al., 2015). Johnston et al. (2017) conducted a visual detection task where participants were asked to view a sequence of five successive images on each trial. The first four images formed a coherent sequence of a transformation, for example, the continuous facial expression. The final fifth image either followed or violated, the expected changes to form the predictable and unpredictable situations, respectively. Their results demonstrated increased N170 amplitude for the unpredictable final image onset than the predictable one. That is, N170 is closely linked with the induced error signals in the brain and may serve as a potential neural index of prediction errors in the predictive coding framework.

1.3.2.3 Accounts for psychophysical phenomena

Predictive coding has also been used to explain various psychophysical phenomena including bistable perception (Weilnhammer et al., 2017; Hohwy et al., 2008), illusory motions (Lotter et al., 2016; Watanabe et al., 2018) and naturalistic sentence comprehension (Shain et al., 2020). In a speech recognition task, Blank et al. (2018) proved that the predictive coding mechanism could offer a more reasonable explanation for the perception/misperception of spoken words than another possible neural mechanism. Participants were required to first read written words as prior expectations and heard subsequent distorted spoken words. Afterwards, they needed to indicate whether the spoken/written word pairs were 'same' or 'different'. For example, the spoken word 'pip' after written 'kip' can be perceived as the same (i.e., 'kip') or different (i.e., 'pip'). The correct perception (i.e. the 'different' response) could be explained by two possible neural mechanisms. In a predictive coding scheme, perception of mismatch is caused by the increase of prediction error signals encoding the deviating sounds (i.e., 'k' and 'p' sounds); Another scheme suggested that it is due to the weaker neural representation of the common sound of word pairs (i.e., the 'ip' sound). Their fMRI analyses suggested that when the written/spoken mismatch was detected behaviorally, neural representations of prediction error were more apparent.

1.3.3 The formation and algorithms of predictive coding

In the current thesis, the concept of 'predictive coding' refers to the computational process defined by Rao and Ballard (1999). To better understand it, here, I give a brief review of the development of this theory as well as relevant extension and variants. A more detailed and comprehensive review of the algorithms of predictive coding can be found in (Spratling, 2017; Millidge et al., 2021).

1.3.3.1 Origination: Linear Predictive Coding in information theory

Information theory (Shannon, 1948) provided inspiration for early neuroscientists who wanted to understand sensory coding and perception. In the field of information theory, the main task is to quantify, compress and transfer digital information. To quantify the information, a key measure is 'entropy', which tells the degree of random variation in the given signal. For example, if there are two audio signals, the first one contains rhythmic sound such as the heart beat sound, the second one contains random white noise. Compared with the second sound, the first one has low entropy, since the sound signal is easy to predict, you would easily figure out when the next beat would occur. To transfer such signal with a repetitive pattern, you may only need to provide the first beat sound and the interval time between two-beat sounds. The second sound is viewed as high entropy since its signal is hard to predict, signals at each time point are unrelated. In order to transfer such sound without losing any information, one needs to record the sound at every time point, which needs higher transmission capacity. Low entropy corresponds to high redundancy and vice versa.

One main problem in information theory is to remove redundancy and therefore only keep the useful messages. This is also important for the brain since the brain is limited in resources and processing ability. 'Linear predictive coding' is a technique used in information theory for transmitting telecommunication signals more efficiently (Harrison, 1952; Makhoul, 1975; Vaseghi, 2000; O'Shaughnessy, 1988). The core idea is that the information at each time point (i.e. the frame in the case of video) could be approximated by a weighted combination of the previous ones. Formally, this could be expressed as:

$$\bar{x}(t) = v_1 x(t-1) + v_2 x(t-2) + \cdots + v_n x(t-n) = \sum_{j=1}^{n} v_j x(t-j) \tag{2}$$

Where $\bar{x}(t)$ represents the estimated frame at time t, n is the number of frames used for the estimation and v is the weight for each previous frame. The idea of 'Linear Predictive Coding' is quite simple and straightforward. However, it is proved to be an efficient way to reduce the bandwidth of message transfer.

1.3.3.2 Introduction: Barlow's minimum redundancy

The introduction of the idea of redundancy reduction to sensory processing in the brain was attributed to the theoretical work of early scientists (Attneave, 1954; Barlow, 1961; Watanabe, 1960). Particularly, Barlow proposed the minimum redundancy principle of neuronal coding and the efficient coding hypothesis. The core idea is that the sensory information is highly redundant in temporal and spatial domains and it will be highly costly for neurons to encode detailed and redundant information. Therefore, we should expect the coding principle should be optimized and the neurons should only fire for the most useful information.

Barlow's formulation relies heavily on Shannon's definition of channel capacity, information, and redundancy (Shannon and Weaver, 1949). Barlow's original definition for redundancy in sensory information could be expressed in Equation (3), where $H(y)$ is the response entropy and C is the channel capacity representing the encoding ability of neurons and determined by physical properties of the encoder. Therefore, to reduce redundancy, the brain should maximize the response entropy through decorrelating the sensory information.

$$R = 1 - \frac{H(y)}{C} \tag{3}$$

In the predictive coding theory (Rao and Ballard, 1999), each level only transmits upwards the unpredictable, error signals instead of the entire representation of the sensory information, which precisely follows the idea of Barlow's minimum redundancy.

1.3.3.3 Predictive Coding Algorithm as formulated by Rao & Ballard.

The mathematical formulation by Rao and Ballard in their seminal work Rao and Ballard (1999) is now considered as the most influential computational process of brain functions.

Model Description.

The predictive coding algorithm formulated by Rao and Ballard was used to implement a hierarchical generative model of visual cortex. However, this idea may be viewed as a general

framework to interpret responses in other brain regions Rao and Ballard (1999). Under the context of visual system, this model could provide inner causes for the external visual stimuli. Inside this hierarchical model, each level could generate predictions or hypothetical causes for the activities in the level below. Consider the ith layer (L_i) in this model, and thus the above layer is L_{i+1} and the below layer is L_{i-1}. A vector $\mathbf{r}$ can denote the neural representation in each layer, with each element in this vector as the activities or firing rates of neurons. If the matrix $\mathbf{W}$ denotes the synaptic weights between two adjacent layers and the function $f(x)$ as the neuronal activation function, then the activities in the layer below can be expressed as:

$$\begin{aligned} \mathbf{r}_i &= f(\mathbf{W}\mathbf{r}_{i+1}) + \epsilon_i \\ \mathbf{r}_{i-1} &= f(\mathbf{W}\mathbf{r}_i) + \epsilon_{i-1} \end{aligned} \tag{4}$$

where $f(\mathbf{W}\mathbf{r}_{i+1})$ and $f(\mathbf{W}\mathbf{r}_i)$ are the predicted activities by the higher level and $\mathbf{r}_i$ and $\mathbf{r}_{i-1}$ are the real responses in the corresponding lower level. Both ϵ_i and ϵ_{i-1} are the stochastic errors representing the differences between the real activation in the lower level and the predicted activation from higher level, i.e. the predictive errors. We can assume that these random errors follow a Gaussian distribution as in $\epsilon_{i-1} \sim \mathcal{N}(0, \sigma_{i-1}^2)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

Optimization

The optimization goal can be twofold: to find the best representation vector $\mathbf{r}$ for given visual stimuli and to learn appropriate synaptic weights matrix $\mathbf{W}$. Interestingly, these two tasks can be achieved in a single optimization or loss function if we take L_i as an example:

$$\begin{aligned} Loss_i &= \frac{1}{\sigma_{i-1}^2}(\mathbf{r}_{i-1} - f(\mathbf{W}\mathbf{r}_i))^T(\mathbf{r}_{i-1} - f(\mathbf{W}\mathbf{r}_i)) + \frac{1}{\sigma_i^2}(\mathbf{r}_i - f(\mathbf{W}\mathbf{r}_{i+1}))^T(\mathbf{r}_i - f(\mathbf{W}\mathbf{r}_{i+1})) \\ &= \frac{1}{\sigma_{i-1}^2}\epsilon_{i-1}^2 + \frac{1}{\sigma_i^2}\epsilon_i^2 \end{aligned} \tag{5}$$

The loss function in terms of L_i consists of two parts, the squared top-down and bottom-up prediction errors, ϵ_i and ϵ_{i-1}. Both squared errors are weighted by the inverse of their

respective variance, meaning the larger the variance, the contribution of that error is smaller.

In order to get the optimal estimation of the representation vector r_i in L_i, which can minimize the prediction errors from both direction, one can perform gradient descent on $Loss_i$ with respect to r_i:

$$\mathbf{r}_i \leftarrow \mathbf{r}_i - \eta \frac{\partial Loss_i}{\partial \mathbf{r}_i} = \mathbf{r}_i - \eta(-\frac{2}{\sigma_{i-1}^2}\mathbf{W}\epsilon_{i-1} + \frac{2}{\sigma_i^2}\epsilon_i) \tag{6}$$

For simplicity, the activation function here is set as an identity matrix, thus $f(x) = x$, η is the learning rate for updating the neural activity.

Comparison with other algorithms

Based on the predictive coding algorithm formulated by Rao and Ballard, Spratling developed another formulation of predictive coding (PC), termed PC/BC-DIM (Spratling, 2008a,b). This version features its compatibility with the Biased Competition (BC) theory (Desimone and Duncan, 1995) of cortical function and its implementation using Divisive Input Modulation (DIM) (Spratling et al., 2009). According to the biased competition theory, external stimuli in the visual field compete for being processed and represented and the result can be biased by multiple mental factors. In PC/BC-DIM, the feedback predictions are treated as the bias for the activity in the lower level, which is realized through the DIM method. Compared with the algorithm in Rao and Ballard where the prediction errors are obtained through subtraction/addition (See Equation (4)), here the errors can be computed by division. Thus, in terms of L_i, the prediction error (ϵ_i) and the neuronal representation ($\mathbf{r}_i$) can be expressed as:

$$\epsilon_i = \mathbf{r}_i \oslash (c_1 + \mathbf{W}\mathbf{r}_{i+1}) \tag{7}$$

$$\mathbf{r}_i = (c_2 + \mathbf{r}_i) \otimes \mathbf{W}\epsilon_i \tag{8}$$

In Equations (7) and (8) $\oslash$ and $\otimes$ denote the element-wise division and multiplication respectively, and c_1 and c_2 are parameters which can prevent $\mathbf{r}_i$ being divided by zero and non-responsive. Spratling's formulation also shows a reasonable implementation of the predictive

coding process. But in the thesis, we will adopt the formula developed by Rao and Ballard because it has been extensively tested in academia.

1.4 The implementation of predictive coding in deep neural networks

In the last section, I discussed predictive coding as a potential unifying theory of brain function. Especially, I described how predictive coding can be formulated and modeled mathematically. In this section, I will introduce a modern technique, deep learning, which has been shown to have more advantages than traditional modeling methods. For instance, deep neural networks can deal with large datasets and show better performance than traditional ones. I will give a brief introduction on deep learning techniques and deep neural networks first and then show how predictive coding can be implemented in a deep learning framework.

1.4.1 Deep learning

As an emerging technology, deep learning is spawned from the interactions between the fields of neuroscience and artificial intelligence (AI) (Yamins and DiCarlo, 2016; Botvinick et al., 2019; Kriegeskorte and Douglas, 2018). In order to understand the meaning of 'deep learning', one has to understand the concepts of 'machine learning' and 'artificial intelligence' first. These three concepts are closely connected to each other and have an inclusive relationship as shown in Figure 10. Artificial intelligence (AI) is a concept opposed to natural intelligence possessed by humans and other animals. It refers to the study of intelligent agents or machines. Machine learning and deep learning thus could be considered as the methods or techniques that help the agents or algorithms to gain intelligence.

Although it is easy to confound deep learning with machine learning, a critical difference should be noted between both concepts. Suppose you want to train an algorithm to classify cats and dogs. Machine learning first needs to define a set of features of cats and dogs such as vectors of ears' shape or body size and then perform the classification based on those predefined feature vectors. Instead, the algorithm in deep learning could automatically learn features or representations from raw data without introducing hand-coded rules like standard

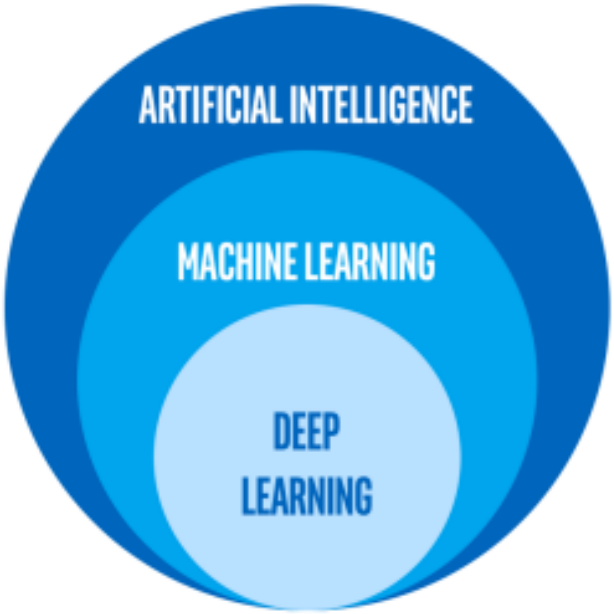


Figure 10: **The relationship between artificial intelligence, machine learning and deep learning.** Artificial intelligence (AI) refers to any non-human agents or machines with intelligence. Machine learning is a sub-field of AI, meaning that the algorithm could automatically improve its performance through learning. As a sub-field of machine learning, deep learning shares the same goal but with less human intervention. Figure adapted from: https://www.intel.ca/content/dam/www/public/us/en/ai/images/ai-machine-learning-deep-learning-rwd.png.rendition.intel.web.480.270.png

machine learning. Therefore, standard machine learning requires more manual human intervention to preprocess the raw data before feeding them to the algorithm. When the algorithm mimics a structure from biological neural networks, it could be termed as an artificial neural network which is a long-standing idea (Rumelhart et al., 1988). Particularly, when the deep learning technique is used, the artificial neural network could also be called a deep neural network. Since the deep neural networks could be enlarged into big sizes, and trained with larger datasets by high-performance graphics processing units (GPUs), they could also be used to solve many new problems, including image classification (Krizhevsky et al., 2012), navigation (Banino et al., 2018) and reasoning (Santoro et al., 2017).

1.4.1.1 How deep learning works?

The implementation of deep learning algorithms or deep neural networks usually involves three steps: Model construction, which illustrates the pathways connections of information flow and also depends on the specific tasks at hand and the dataset to be processed; Model training or learning, a step to update the parameters in the model in order to fit the training dataset; Model testing, which evaluates whether the model will perform well with previously unseen data.

Model Construction

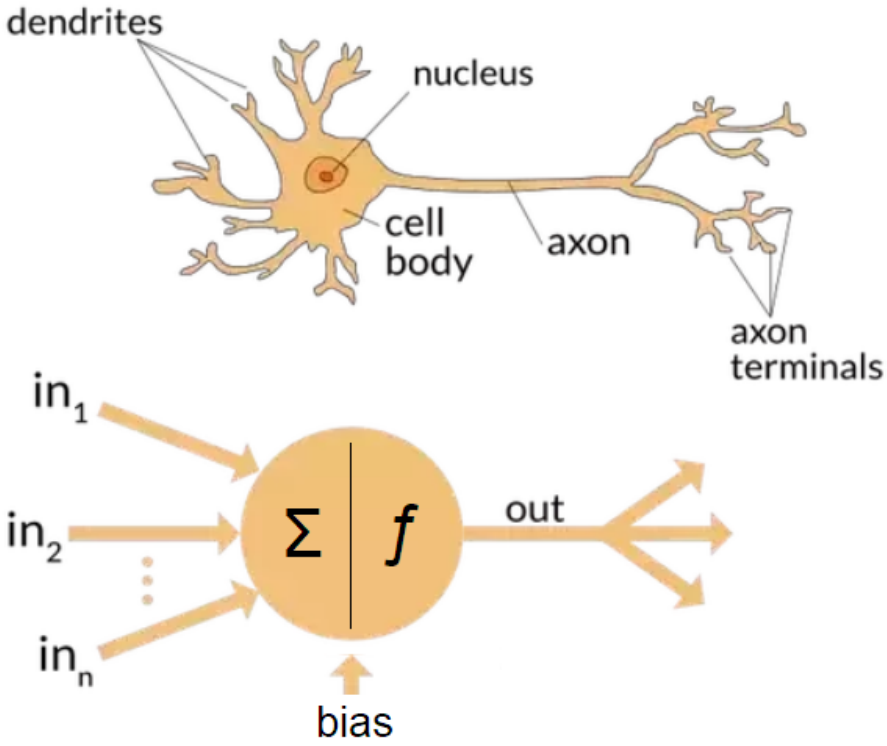


Figure 11: **A biological and an artificial neuron.** Figure adapted from the link (https://www.quora.com/What-is-the-differences-between-artificial-neural-network-computer-science-and-biological-neural-network)

The construction of an artificial neural network (ANN) involves the arrangement and connections of artificial neurons. Just like biological neurons act as the basic structural and

functional units of the nervous system, the artificial neurons are the smallest constituent units constituting the architecture of ANNs. As shown in Figure 11, multiple input values are added up or integrated inside the artificial unit. The integrated information is then biased and activated by certain rules or functions to form output values. That is, the artificial neurons loosely mimic the integration and activation properties of biological ones (Rumelhart et al., 1988).

The architecture of a deep neural network determines the organization of artificial neurons as well as the flow of information (Richards et al., 2019). Architectures can vary largely depending on the specific tasks and desired functions performed by the networks. Usually, a deep model could include an input layer, output layer and hidden layer(s) meaning all the middle layers, as shown in Figure 12. This is where 'deep' comes from in deep learning. A 'layer' consists of units or artificial neurons arranged in a certain way and should be considered as being analogous to brain areas instead of the laminar structure in cortex architecture (Richards et al., 2019). Particularly, the nonlinear output and multi-layer architecture endow the deep network with a hierarchy of increasing complexity and abstraction which is similar to the hierarchy in the brain cortex.

Usually, units in the same layer stay unconnected, while units in the different layers will connect in various ways. According to the connection patterns, ANNs can be divided into different categories. For example, a multilayer perceptron (MLP) has all the neurons fully-connected; while Convolutional neural networks (CNNs) can have their neurons partially connected (I will go into depth for this part in the next section). Since connections between neurons can store information of weights, more connections mean more parameters contained by the network. Thus, the fully-connected model will include much more parameters than CNNs, meaning a higher demand for computational ability of the computers.

Model Training

If one wants to use an artificial neural network to perform a specific task, like image classification, it is not enough to only have its components connected. The model needs to

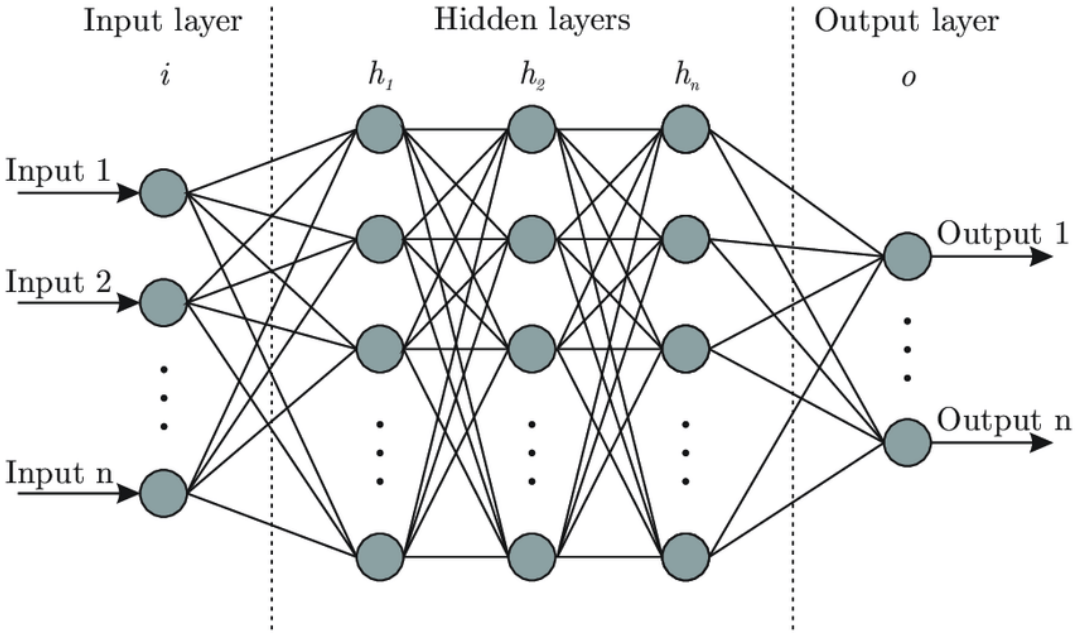


Figure 12: **A typical arrangement of deep neural network.** It usually contains three parts: Input layer, output layer and one or multiple hidden layer(s). Here, neurons are fully connected with each other. Figure adapted from (Bre et al., 2018)

be trained to obtain high performance since the connection parameters are random at the beginning. Essentially, training a model means assigning appropriate weight parameters for the connection of the model. Suppose you want to train a model to classify pictures of cats and dogs. The model needs to 'see' (i.e. process through initial random weights) pictures of cats and dogs (i.e. the learning material or training dataset) and then judge their classes. The decisions of the model are compared with the ground truth of the pictures (i.e. their labels). The resulting mismatches between predicted classes by the model and the real ones will be computed and used for adjusting the parameters' value for better prediction next time. The above steps will be iterated until the model can predict the class of a given picture with high accuracy, i.e. the parameters are adjusted to appropriate values.

Traning Dataset. There are various datasets. The choice of dataset depends on the specific task at hand. If one wants to get a model that can classify cats and dogs, then a dataset including a lot of cat and dog pictures will be needed. ANNs can be used for many tasks, such as natural language processing, translation, self-driving, classification, etc. In the field

of computer vision, the commonly used datasets include ImageNet (Deng et al., 2009), Cifar10 and Cifar 100 (Krizhevsky et al., 2009), MNIST (Deng, 2012) etc. Figure 13 shows some training samples for handwriting recognition.

Figure 13: **Handwritten numbers.**

Objective Function. The way to quantify the above-mentioned mismatches between model predictions and ground truth is to formulate an objective function, which can also be referred to as a loss function. It tells the distance or loss between the current model performance with a suboptimal set of parameters and the best one with optimal parameter configuration. Therefore it can serve as criteria describing how 'good' or 'bad' the current model is. Figure 14A considers an extremely simplified situation where the model only contains one parameter. For a given parameter value, the red dot, the function can indicate the corresponding performance of the model. In fact, the objective function provides an overview of the model

performance in the entire parameter space.

Recall the nature of model training or learning is to obtain the optimal set of parameters. Graphically, the model will perform best when the red dot sits at the lowest location of the objective function in Figure 14A. This can be achieved by 'learning rules' as shown in Figure 14B where only two parameters are contained in the model. The learning rules indicated the direction and speed of updating the model's parameters.

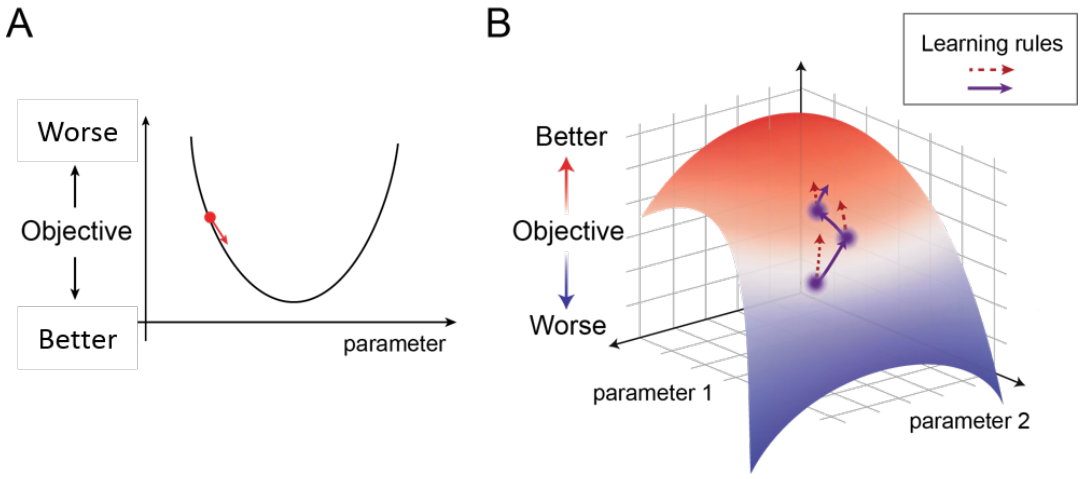


Figure 14: **Objective function.** Figure B is adapted from (Richards et al., 2019)

Supervised and Unsupervised learning. The model can be trained by different approaches. Generally, the model can be trained in a supervised or unsupervised way. The former refers to learning with an ground truth. For example, when feeding a dog picture to the network during training, the network needs to predict whether the presented picture is a dog or a cat, and then the predicted answer (dog) needs to be compared to the real answer (cat). If inconsistency occurs, a big loss will be generated by the objective function to force the updating of the weights for a better prediction next time. However, unsupervised learning doesn't need an answer. You only need to provide pictures without needing to know what the picture is. In this situation, the objective of the network is to represent the input images as accurately as possible. It has been said that unsupervised learning is more in line with the

learning of humans, since most of the time, we can represent and store an object as accurately as possible without knowing its name. Therefore, supervised learning is task-driven; while unsupervised is data-driven.

Model Testing

After the model is well trained on a training dataset, one cannot use it directly in tasks. For example, a model may be trained to perform classification with high accuracy on the training dataset. However, this may not be true for the new, unseen data. Therefore, the model needs to have good generalization. One way is to use a validation set as a reference as shown in Figure 15. The validation set is not designed for training the model but to evaluate the performance of the model. Instead of training the model for lower training loss, one has to have an early stop where the validation loss is lowest as indicated by the red vertical line.

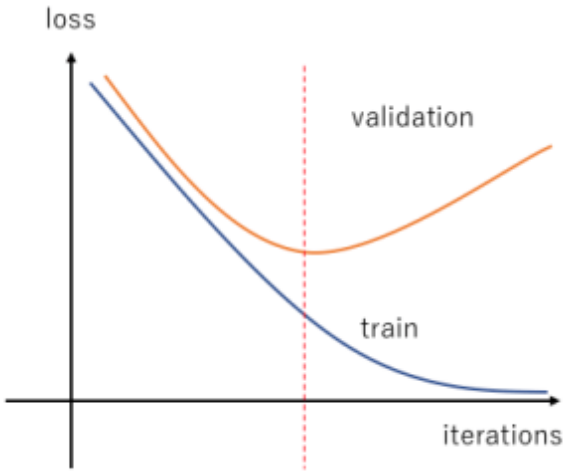


Figure 15: **Model testing.** Figure from the link (https://larevueia.fr/wp-content/uploads/2020/09/loss-curve-overfitting.png)

The above shows some basic knowledge about how to build, train and test artificial neural networks. In the next section, I will introduce different types of deep neural networks including models used in the current thesis.

1.4.1.2 Types of deep neural networks

Feedforward vs. recurrent neural nets

According to the connection patterns between artificial units, the signal flow will be different. The simplest way of connecting units is in the feedforward direction, meaning information goes directly from the input layer to the output layer. Historically, the feedforward neural network was the first type of artificial neural network. The simplest feedforward neural network is a single-layer perceptron network, which included the perceptron computational units (McCulloch and Pitts, 1943; Rosenblatt, 1961).

Different from feedforward neural networks, recurrent neural networks (RNNs) adopt another way of connecting model nodes. Inside RNNs, the connections between nodes can be directed or undirected graphs along a temporal sequence, which endows RNNs with temporal dynamics. The most significant features of RNNs are that they can use their internal state retained from the last temporal step and process input data with variable length, which is distinct from feedforward neural networks.

Autoencoder

An autoencoder is also a type of artificial neural network which can process unlabeled data. Therefore it can perform unsupervised learning during training. The autoencoder could be divided into two parts: one part can encode input into representational code; another part can decode that code into reconstruction data. It was originally used as a way to reduce dimensionality or learn features and now it was widely used for generative models in the field of machine learning or deep learning. The simplest pattern of an autoencoder can be formed by an input layer, the hidden code, and an output layer.

Convolutional neural networks

A Convolutional Neural Network (CNN) is a deep learning algorithm used primarily in the field of computer vision. Its generation is out of the intention of enabling machines to view the world as humans do. The emergence of CNN was inspired by the observation of the ventral visual pathway (Fukushima and Miyake, 1982). Through mimicking the

hierarchical architecture and the growing receptive field property, the CNNs could replicate the basic trade-off between features selectivity and position invariance in the ventral pathway.

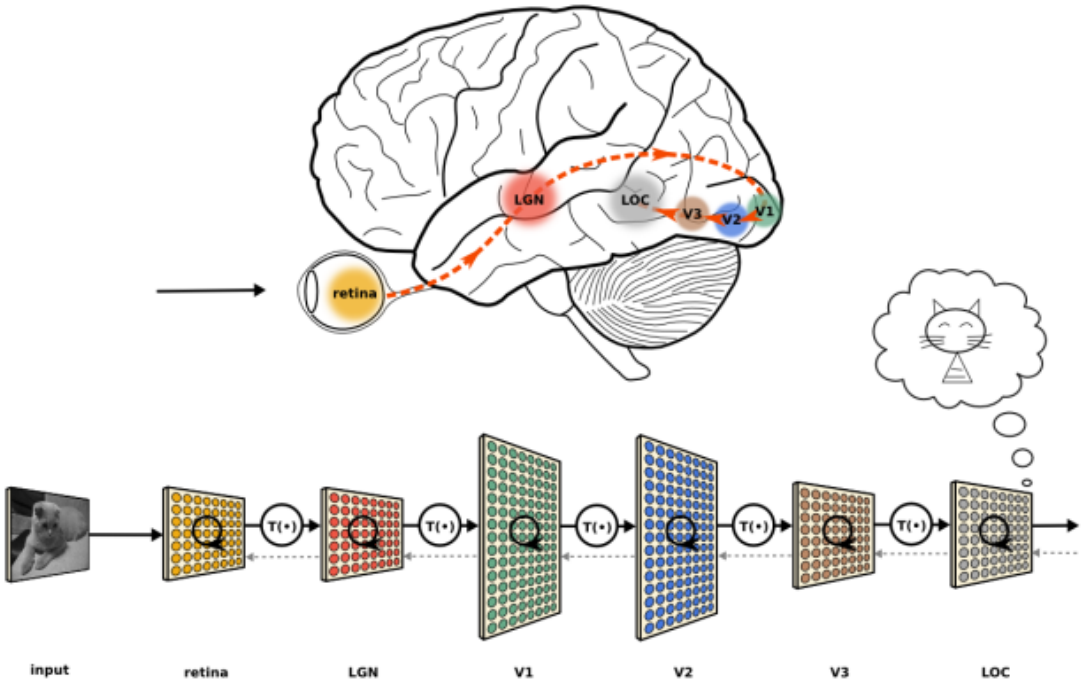


Figure 16: **The correspondence between biological brains and CNNs.** Figure adapted from: https://neuwritesd.files.wordpress.com/2015/10/visual_stream_small.png.

An evident advantage of CNNs is its property of position invariance. Consider a 3×3 basic binary image. In order to recognize the content of the image, early vision models like Multilayer Perceptron (MLP) (Haykin, 1994) will first flatten the image into a 9×1 vector. However, this method will lose the spatial information in the image. For example, when a dog is presented in the center of the picture, the network may recognize it like a dog; while when the dog was shifted to the corner of the picture, the network may probably fail to detect it. Compared to biological vision, we can recognize a dog wherever it appears. To solve this problem, the neural network needs to gain the position/translation invariance in terms of visual objects as observed in biological vision. (LeCun et al., 1998).

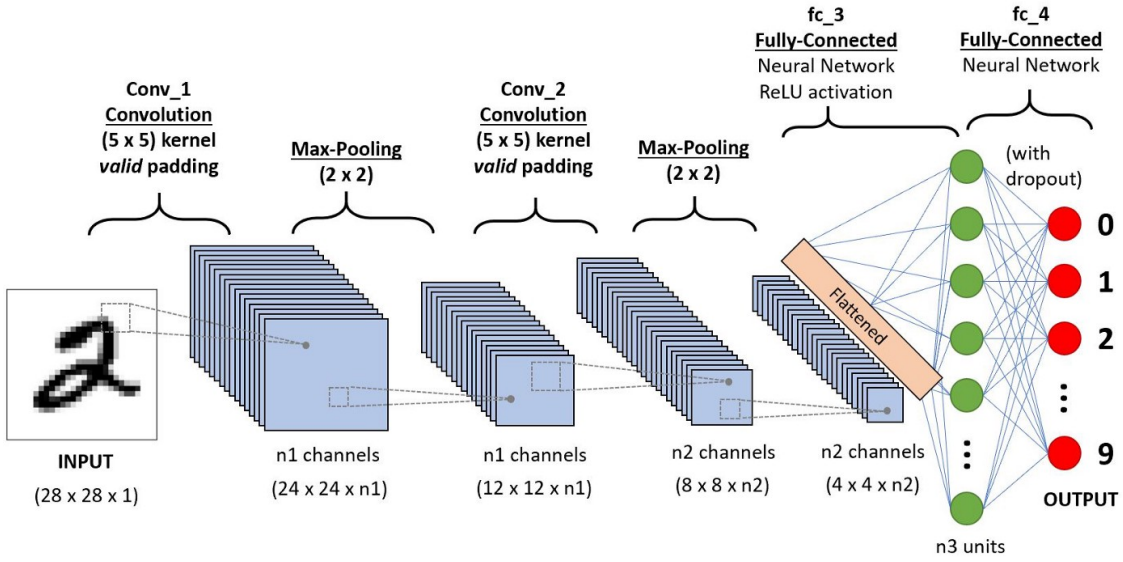


Figure 17: **A CNN architecture for classifying handwritten digits.** Figure adapted from: https://miro.medium.com/max/1400/1*uAeANQIOQPqWZnnuH-VEyw.jpeg

1.4.1.3 Training of deep neural networks

Training models in deep learning is an important step, which can directly affect the performance of the models of interest. The following shows several critical concepts related to the training of models.

Gradient descent

The gradient descent (GD) method is a first-order optimization algorithm. It is important and basic in the fields of machine learning and deep learning since it could be used to find a local minimum or maximum of a predefined objective function that measures the performance of the network on a dataset compared with the ground-truth target label.

Backpropagation

Backpropagation means "backward propagation of errors", which is used for supervised learning of deep neural networks. When using a supervised learning to train a model, errors between data class and the predicted class by neural networks will be calculated. The algorithm of backpropagation could calculate the gradient of error function with respect to

the neural networks's weights to optimize the weights in deep neural networks. The concept of backpropagation contrasts with the naive approach of calculating the gradient of each layer separately in a forward flow. Backpropagation is a more efficient algorithm than the traditional one since the backward transmission of gradient can reuse the already calculated gradients, which can greatly increase the computational power.

Backpropagation through time

Backpropagation through time (BPTT) is a training algorithm for recurrent neural networks. In order to update the weights in RNNs, BPTT could unroll the network over time. That is, at each timestep, one can get one copy of the network with input and output. The resultant errors could be calculated and summed up for all timesteps.

1.4.2 Predictive coding in a deep learning framework

Now we have gained basic knowledge about the deep learning technique and deep neural networks. Generally speaking, compared with traditional implementations of predictive coding, i.e., implementing the direct mathematical formulations, the usage of a deep learning framework can show several advantages. For instance, deep predictive coding models can scale to a very large architecture but stay efficient. In recent years, few predictive coding algorithms have been implemented with deep learning techniques, attempting to improve object recognition. In this subsection, I will review these studies.

PredNet. The earliest deep predictive coding network may be the PredNet (Lotter et al., 2016) which was used to predict the next frame of a video sequence. In terms of model training, they employed an unsupervised objective, which has been shown to be closer to human learning in reality than standard supervised learning methods. Because most of the time, human learning involves accurately describing what they see, instead of being told what it is by an instructor. Nevertheless, it bears mentioning that the ability of their model to predict the next frame in the video stream may mainly be because that it was trained to predict.

PCNs. Predictive Coding Networks (PCNs) were used to study how predictive coding helps improve the classification of clean images (Wen et al., 2018; Han et al., 2018). Their equations are very similar to those proposed by Rao and Ballard (1999), which makes their models competent for adapting the algorithm of Rao and Ballard in deep learning frameworks. However, an obvious violation of Rao and Ballard's scheme should be noticed. Instead of minimizing reconstruction error, their model minimizes classification error. That is, they employed a supervised learning method, which does not fulfil the predictive coding objective as described by Rao and Ballard (1999).

SDPC. 'Sparse Deep Predictive Coding' (SDPC) model constructed by Boutin et al. (2021) is a predictive coding network with a sparsity constraint. The model simulated both the intrinsic structure within each level of the early visual system by the sparse coding component and the extrinsic reciprocal connectivity between layers through the predictive coding component. Their results showed the model could learn similar receptive fields to neuron in V1 and V2 and also provided evidence that the feedback helps the network perform contour integration. Moreover, their model showed robustness to noisy inputs in the context of reconstruction.

Predify. The implementation of Predify by (Choksi et al., 2021) shows some similarities with PCNs. Both networks were designed under a similar goal of improving object recognition. The differences are also evident. In terms of model learning or weights updating, Wen et al. (2018) performed the optimization only with a classification objective, which leads to non-uniform reduction of Reconstruction errors over timesteps. However, the Predify model was optimized with both classification and reconstruction objectives, which causes an increasingly better performance over timesteps, which is more biologically plausible. In the current thesis, we adapted the model described by (Choksi et al., 2021).

In summary, although these models may differ in tasks, objective functions, and training methods, their results show that predictive coding can help improve the model's performance. This suggests that a deep model with predictive coding may perform better than other models or even have the potential to show human-like performance if we carefully

design and implement the model.

1.5 Brain oscillations: neuronal bases of predictive coding?

If predictive coding could act as a unifying theory to explain most of the phenomena of brain function. Which brain activity in the brain could be used to implement such a strategy? It seems that the most significant activities that happen in the brain are oscillations (or called brain waves). They are ubiquitous in the brain and are thought to be closely related to various brain functions. Is it possible that the predictive coding strategy is implemented by virtue of oscillations? In other words, oscillations could convey or represent prediction error signals? To answer this question, in this section. I will first present a brief introduction to brain oscillations in terms of their properties. Particularly, I will point out a special property of oscillations–propagation through cortical regions, thus gaining the name "traveling waves". Lastly, I will give some evidence showing how traveling waves could be related to the predictive coding mechanisms.

1.5.1 Brain oscillation

1.5.1.1 What are brain oscillations

Brain oscillations or brainwaves refer to the rhythmic neural activity in the brain cortex. The first human oscillatory activities were recorded by Hans Berger in 1924 through electroencephalography (EEG), which is a device also invented by Berger. Therefore he is commonly accepted as the forefather of EEG. EEG can obtain electrical activity or electroencephalogram on the scalp. In his seminal paper, Berger (1929) described two distinct recorded signals which he termed as alpha rhythm and beta rhythm. The former one is a relatively slow oscillation. They were detected from the occipital regions. When subjects stayed in a relaxed wakeful state, the rhythm presented and when the eyes are closed, the rhythm increased. The latter one, a relatively fast oscillation appeared with opened eyes.

1.5.1.2 Measurement of EEG signals

The oscillatory activities in the brain arise from the electrical activities which are widely observed at different levels of brain organization or structure. According to Haken (1996), three levels can be considered: (i) the microscopic level that concerns the activity of a single neuron; (ii) the mesoscopic activity of a local group of neurons; (iii) the macroscopic activity involving different brain regions. Activities at different levels are entitled to different names and measured with different devices. However, in nature, they are rhythmic electrical activity.

At microscale. Single neurons could generate an action potential or spike when the electric membrane potential reaches the critical threshold, which starts from the neuron's cell body, through the axon and finally reaches to the next neuron forming post-synaptic electrical signals. The rhythmic firing of action potentials could form oscillatory patterns. The subthreshold fluctuation of membrane potentials, i.e., without resulting in action potentials, could also be oscillatory. Therefore both types of activities could form the oscillations of membrane potential. The activities of a single neuron are considered fundamental for information transmission in the brain. However, they are way too small to be detected outside the scalp. Usually, the invasive measurement of single-unit recording will be employed to monitor a single neuron's activity. Such measurement is usually performed on non-human animals, such as mice.

At mesoscale. This level records the activities of numerous neurons or neuron populations. That is, multiple neurons need to spike or fire in synchrony, oscillations at the population level can be detected, otherwise only noise can be observed. Mesoscopic oscillations can be termed as local field potentials (LFPs). Although their activity is stronger than that of single neurons, to obtain clean data, intracranial electrodes are required. Such experiments are usually performed on patients with a brain disease like stroke.

At macroscale. Neural activity generated by large groups of neurons is widely studied since they can be measured outside the scalp using techniques like electroencephalography (EEG)

and magnetoencephalography (MEG). Subjects can be measured in a non-invasive way, which is an advantage for data collection. Usually, EEG recordings reflect the synchronous activity of thousands or millions of neurons, which renders them not as accurate as LFPs or single-unit recordings. Importantly, the oscillatory pattern of large-scale activity does not need to match the firing pattern of individual neurons, which leads to the interaction between the two activities which will be detailed in the following section.

The current thesis concerns how large-scale oscillations play roles in various cognitive functions and how they could be related to prediction coding processing. Therefore, the following sections will focus on large-scale oscillations.

1.5.1.3 Characteristics of oscillations

The recorded brain oscillations could be described by three characteristics: frequency, amplitude, and phase. Frequency describes how fast the oscillations are and they have unit 'Hertz' (Hz), meaning how many cycles are shown in one second. For example, 10 Hz means 10 cycles in one second. The higher the number of cycles, the higher the frequency or the faster the oscillation. Amplitude denotes how strong the oscillations are (see Figure 18). The unit could be magnitude or power (square of magnitude). Phase means the position of the oscillations (in radians and degrees). These three properties are dependent on each other. For instance, when the amplitude of a given oscillation becomes zero, the other two properties also disappear, meaning no oscillation. Also, it has been shown that most biological signals follow a '1/f' power profile, meaning higher frequencies have lower amplitude. This is also true for brain EEG signals (Roopun et al., 2008).

The frequency of EEG oscillations could range from 1 to about 100 Hz. Traditionally, these oscillations could be grouped in frequency bands: delta (or δ, 2-4Hz), alpha (or α,7-13Hz), beta (or β,13-30Hz) and gamma (or γ,30-100Hz). Later work tried to uncover whether there exists a functional separation between frequencies bands (Buzsáki and Draguhn, 2004; Penttonen and Buzsáki, 2003; Roopun et al., 2008). The computational results from Penttonen

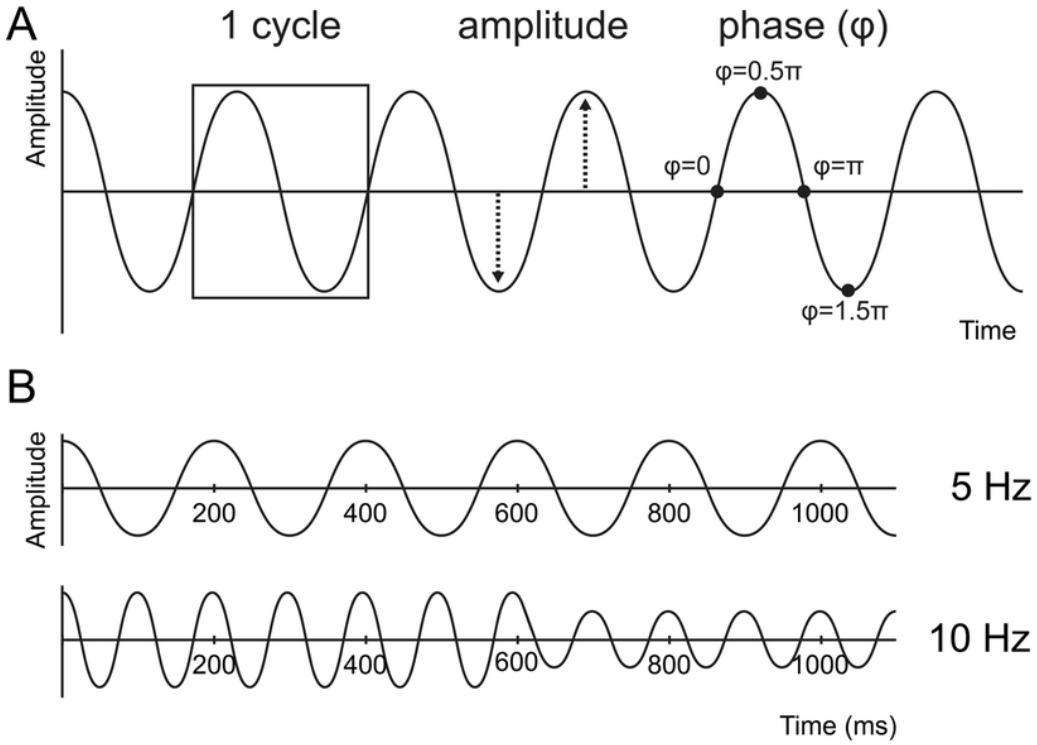


Figure 18: **Characteristics of brain oscillations.** A. Frequency means how many cycles in one second; amplitude means the distance of a peak or trough deviating from the x-axis; phase indicate the position of a point on the oscillation. B. Two types of oscillations with frequency 5 Hz and 10 Hz. Figure from (van de Vijver and Cohen, 2019).

and Buzsáki (2003) suggested at least 10 independent functional frequency bands, with 5 falling within the above-mentioned traditional bands.

A wide range of studies has shown that the specific frequency band seems to be linked with different cognition processes. Generally speaking, slower oscillations such as delta, theta, and alpha could synchronize cortical activity over a large spatial scale (Buzsáki and Draguhn, 2004); while faster oscillations such as beta and gamma are more likely to play a role in relatively smaller brain areas (Sauseng and Klimesch, 2008). Specifically, delta oscillations could be involved in language processes (Sauseng and Klimesch, 2008). Theta oscillations have been reported to play roles in memory (Lisman and Idiart, 1995) and executive functions (Cohen and Donner, 2013; Gulbinaite et al., 2014). Alpha oscillations could be linked with various cognitive functions, including inhibition (Jensen and Mazaheri, 2010) and spatial at-

tention (Foxe and Snyder, 2011). Beta oscillations are linked with somatosensory functions (Baumgarten et al., 2015; Neuper and Pfurtscheller, 2001) and finally, gamma oscillations, as an extra-fast frequency band, are also reported to underlie a wide range of cognitive functions, including neuronal communication (Fries, 2005) and visual awareness (Engel and Singer, 2001).

1.5.1.4 Types of brain oscillations

In addition to being sorted by frequency, brain oscillations can also be classified based on their relation to external stimulation.

Ongoing responses

Ongoing responses refer to activities or oscillations generated spontaneously or endogenously instead of being caused by external stimulation. It is believed that the spontaneous or ongoing oscillations can reflect the brain excitability (Steriade et al., 1993). Even if they are not directly correlated to the information processing, their states can somehow determine the visual outcome; and in turn, the visual stimulation can also modify their states (Nunez and Srinivasan, 2006).

Evoked responses

Compared to ongoing responses, the evoked responses can refer to brain activities that are phase- or time-locked to the presentation of stimuli such as a light flash or a pure tone. In other words, it reflects a resetting of the ongoing oscillation. Usually, the evoked potential amplitudes tend to be low, especially due to the existence of ubiquitous unrelated spontaneous oscillations in the brain as well as the ambient noise. To resolve this, signal averaging can be performed. This is due to the unrelated noise occurring randomly while evoked responses are locked the stimulus, which allows the undesired noise to be averaged out in the time domain (Misulis and Fakhoury, 2011). In practice, to get a discernible evoked response, experimentalists need to obtain repeated trials for the averaging procedure.

One example of evoked response can be the visual evoked potential (VEP). As the name suggested, it is typically caused by a visual stimulus. Thus, the evoked responses usually originate from the occipital visual areas. The VEP can measure the temporal duration for a visual stimulus to pass from the retina to the occipital cortex. This approach has been extensively applied to study attention (Luck, 2014; Luck et al., 2000; Woodman, 2010) and sensory processes (Hillyard et al., 1998), although it is still unclear whether the VEP is caused by a resetting effect or an additional signal superposition effect of visual stimuli on the background ongoing oscillations (Becker et al., 2008; Makeig et al., 2002; Mazaheri and Jensen, 2010).

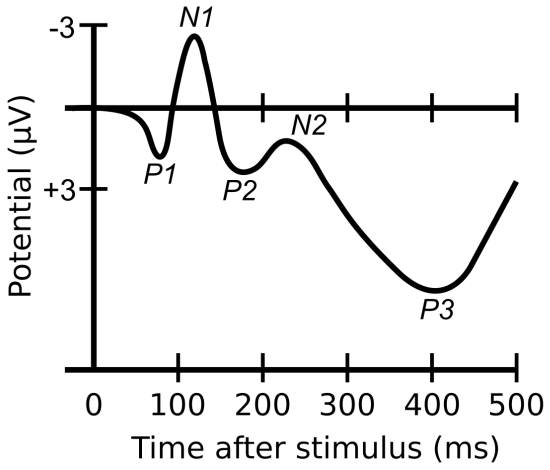


Figure 19: **The time course of ERPs.** Figure from wikipedia.

Another example can be event-related potentials (ERPs). Compared to VEP, they have longer latency and are supposed to reflect the process of higher cognition. Compared with pure EEG data, ERPs can assess the highly specific neural process underlying different sensory, cognitive, and motor events. As shown in Figure 19, the temporal profile of ERPs consist of a sequence of positive and negative voltage fluctuations, reflecting underlying processing components (Luck and Kappenman, 2012). Most ERPs components are named after their polarity (with N/P denoting negative/positive waveform) and latency in milliseconds (with 100 indicating 100 milliseconds after the onset of the stimulus) or its ordinal position (with

1 in 'N1' meaning the first negative peak) in the waveform. The timing of the ERPs components is thought of as an indicator of the timing of the underlying information processing, which offers the potential for revealing the physiological correlates of cognitive processing of interest. For instance, N170 has been reported to reflect the neural processing of faces, familiar objects, or words (Rossion et al., 2003; Hillyard et al., 1998).

In summary, the evoked responses can reflect the underlying oscillatory components that are related to the presentation of a particular stimulus or the processing of that stimulus.

Induced responses

Like evoked responses, the induced ones are also caused by stimulation, yet the latter are not time-locked, meaning their latency can vary through trials (Tallon-Baudry and Bertrand, 1999). Therefore, the averaging technique used in evoked response is not applicable here, for it will cancel out the induced responses through trials. Although one cannot inspect the time course of an induced response like in evoked responses due to their fixed latency, the content inside induced responses can be stable and closely related to the inducing target. One way to obtain the content is using time-frequency transforms, which can reveal the related frequency band to the inducing event. It has been reported that induced oscillations have been linked with various cognitive functions. For instance, the induced oscillations in gamma band are found to relate to object representations (Tallon-Baudry et al., 1996), awareness (Wyart and Tallon-Baudry, 2008) and sensory processing (Baldauf and Desimone, 2014).

1.5.2 The functions of brain oscillations

Oscillations are ubiquitous in the brain and they are the most prominent components of brain dynamics. Initially, brain oscillations are considered by-products, as hypothesized in conventional neurophysiological models (Shadlen and Newsome, 1998). But it is now evident that the functions of brain oscillations are extensive. Their functions can be roughly summarized into two levels: (i) At the neuronal level, how large-scale background oscillations

coordinate activities of individual neurons as well as influence the excitability of cortex; (ii) At the cognitive level, how oscillations influence or modulate various cognition functions such as attention, consciousness, and perception, etc.

1.5.2.1 Oscillations & Neurophysiological responses

Individual neurons can exchange information by sending spikes between each other. It has been proposed that background oscillations (i.e., the synchronized oscillatory activities of a large number of neurons) may underlie effective and flexible neuronal communication, which in turn is essential for various brain functions at a cognitive level. The Communication-through-coherence (CTC) hypothesis (Fries, 2005) provides a possible scheme for how oscillations play their role in neuronal communication. It states that the excitability of neuron populations could be modulated by the phase of oscillations (see Figure 20). Therefore, at a particular phase moment, individual neurons are allowed to send and receive messages efficiently.

It has been hypothesized that sensory processing is closely linked with brain oscillations. Especially, brain oscillations might serve as a carrier for sensory messages. Many studies examined which characteristics of oscillations are crucial for encoding signals. For example, Kayser et al. (2009) quantitatively demonstrated that more information can be encoded in phase instead of the power of neural oscillations. This phase coding of oscillations seems consistent with the results in individual neurons. How do individual neurons encode information? It has been hypothesized that neuronal spikes may carry information with two approaches. The first one is a traditional coding scheme called rate coding or frequency coding, stating that the frequency of firing rate of neurons is positively related to the stimulus' intensity (Kandel et al., 1991; Adrian and Zotterman, 1926). However, this regime excludes the possibility that information may be encoded in the temporal structure of the spike train, which makes the theory too simplistic to describe the brain encoding (Stein et al., 2005). The second is a temporal coding approach, meaning the precise spike timing can carry information

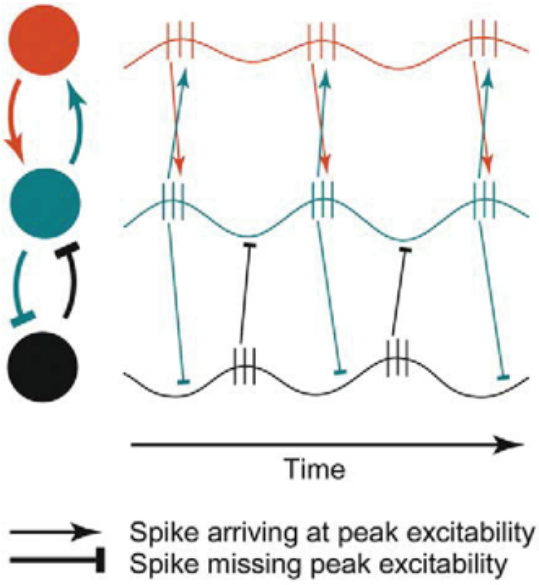


Figure 20: **Communication through coherence.** Three individual neurons are modulated by different oscillatory patterns. The oscillations for red and green cells fluctuate with the same phase position, leading to efficient communication between them. However, the oscillations for green and black cells show different phase pattern, which blocks the signal transmission. Figure adapted from (Fries, 2005).

(Gerstner, 2002; Dayan and Abbott, 2001). Compared to frequency coding, temporal coding is more efficient.

1.5.2.2 Oscillations & Attention

Attention is one of the most important brain functions, which can allocate and concentrate brain resources on a particular task or event. For instance, when you are focusing on video games without noticing the people passing by, this is because attention concentrates the brain resources on that game. Suddenly you hear someone calling your name outside the window and then you stop playing and look for the source of the sound. That's the allocation/shift of attention to another event. Researchers often use 'Spotlight' as a metaphor for attention, meaning the concentration of limited brain resources on a particular target surrounded by tons of unrelated information. Attention can operate in either a 'bottom-up' or 'top-down'

approach. The former involves an involuntary drive by external salient stimulation, while the latter describes a voluntary process of inner resources' allocation by attention. It has been reported that when an external or internal event is under the spotlight of attention, the corresponding process for that event will be enhanced (Moran and Desimone, 1985).

There is a long-standing debate on the nature of attention: is attention continuous or discrete? Previous studies held a continuous idea that attention is indivisible (Posner and Petersen, 1990; Treisman and Gelade, 1980). However, recent studies revealed the discrete and rhythmic nature of attention processing (VanRullen and Dubois, 2011; Buschman and Miller, 2007; Wolfe et al., 2011).VanRullen et al. (2007) performed a psychophysics study and reported that attention can sample multiple items in the visual field sequentially. It is worth noting that the sampling rate of attention is at theta range. In another demanding visual search task, Dugué et al. (2015) presented similar results, showing that spatial attention is distributed to each of the visual stimuli periodically at ∼7Hz.

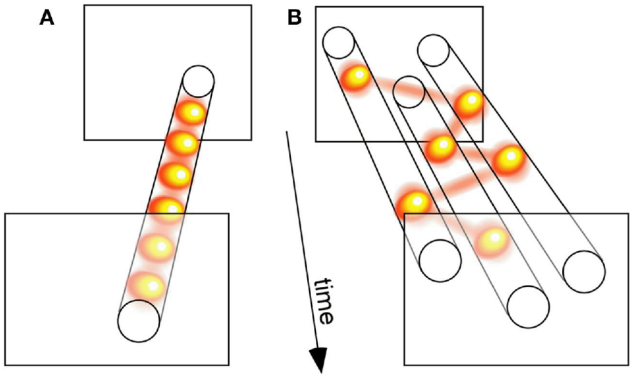


Figure 21: **Attentional rhythms.** A. When one location/visual input is attended, attention samples it periodically at ∼7Hz. The fireballs represent the moments when attention can sample, which is modulated by the underlying neural process. B. When attending multiple objects, attention samples them successively, causing each object is sampled by fewer times, corresponding to low sampling rate. Figure adapted from (VanRullen and Dubois, 2011)

The frequency of the rhythm of attentional sampling can vary under different task conditions. In behavior tasks, subjects are required to attend to one or two visual objects. Their attention

is reset by a cue for aligning the phase of attentional rhythms. Two studies showed that attention samples each of the two visual items sequentially at 4Hz (Fiebelkorn et al., 2013; Landau and Fries, 2012). Does attention sample with different frequencies? It is maybe easy to reconcile these discrepant experimental results. Attention can sample at a fixed internal frequency, ∼7Hz. As shown in Figure 21, when attending more than one visual object, attention samples them alternatively, resulting in lower frequency at each sampling location (VanRullen and Dubois, 2011; VanRullen, 2016). Supporting evidence for this view came from a study by Holcombe and Chen (2013) which reported a 7Hz sampling rate for tracking one object and ∼3Hz for three objects.

What is the neural mechanism underpinning attentional sampling? It has been proposed that brain oscillations may relate to attentional sampling, especially theta-band oscillations. Landau et al. (2015) revealed the modulation of 4Hz oscillation for gamma-band activity when subjects attend two objects. Since gamma band frequency can be indicative of information processing, the 4Hz brain oscillations will directly relate to the observed attentional rhythms.

1.5.2.3 Oscillations & Consciousness

Consciousness can have multiple meanings. Here I restrict consciousness to the meaning of conscious perception which can be measured by the 'reportability' of an item, as put forward by Dehaene and Changeux (2011), 'if you are conscious of something, you are able to report it.' One important topic in the field of consciousness is to find its neural correlates. That is, how consciousness is generated by the physical brain? A widely accepted strategy for this issue is contrastive analysis (Baars, 1993). The key is to create two conditions with one involving conscious processing while the other does not. The difference between the two corresponding brain states may provide insights into the underlying neural correlates.

Could brain oscillations serve as the neural correlates of consciousness? By presenting a near-threshold stimulus, it is possible to link the awareness for the visual stimuli and corresponding oscillatory neural activities. This will be reviewed in the next section. Inversely, the state of

consciousness can also affect oscillations. Compared to the unconscious state, consciousness promotes the synchronization of oscillations, meaning the communication between brain regions. For instance, the conscious perception of words results in gamma (Melloni et al., 2007) and beta (Gaillard et al., 2009) synchronization in a long-range. Except for conscious perception of words, when words are retained in memory, an even stronger conscious state, theta activity in the frontal areas is enhanced. In summary, brain oscillation may serve a causal role for consciousness; in turn, consciousness can increase the activity of oscillations as well as their synchronization across multiple areas.

1.5.2.4 Oscillations & Perception

Perception can be viewed as the interpretation of sensory information. Is our perception continuous or discrete? Most of the time, our perception seems to be smooth and continuous. However, suppose you are watching fast turning wheels under steady illumination. The continuous hypothesis cannot explain the resulting wagon-wheel effect meaning the wheel that turns forward is perceived as turning backward. Instead of directly considering perception as discrete, VanRullen (2016) proposed the concept of 'rhythmic perception' (see Figure 22). Specifically, the ability of perception can fluctuate, with strong perceptual ability at some moment and weak at another moment, which produces 'perceptual cycles'

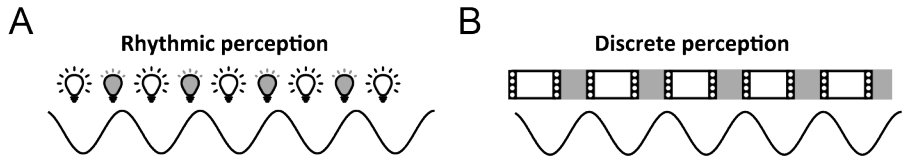


Figure 22: **Perceptual cycles.** Adapted from (VanRullen, 2016)

Many studies have demonstrated that perceptual cycles can directly link with the brain oscillations in theta and alpha bands VanRullen (2016). The moment of strong perceptual ability can be viewed as high cortical excitability. Bishop (1932) first demonstrated that the LFP phase is closely linked with cortical excitability.

1.5.3 The propagating oscillations: Traveling waves

Brain oscillations can provide a very fine time course of sensory processing, which is closely related to various cognitive processes. In the spatial domain, an increasing number of studies suggest that these oscillations could be organized as traveling waves across brain regions. Interestingly, in the predictive coding framework, the transmission of signals (prediction errors and prediction) is transmitted between different hierarchical levels, i.e. brain regions. Is it possible that the traveling waves with certain directionality can carry signals postulated by the predictive coding theory? In this subsection, I will give a brief introduction about traveling waves, and the in next subsection, I will discuss the possibility of traveling waves serving as neural mechanisms of predictive coding.

1.5.3.1 What is traveling waves

Traveling waves can refer to any traveling brain activities, but in this thesis, I will constrain its definition as periodic or oscillatory traveling waves. That is, here, the concept of traveling waves is considered the spatial property of brain oscillations across brain areas. The existence of traveling waves has been reported across multiple species (Ermentrout and Kleinfeld, 2001; Sato et al., 2012), at differing scales of measurements (Muller et al., 2018), and under various stimulation conditions (Nauhaus et al., 2012; Sato et al., 2012).

1.5.3.2 The functional role of traveling waves

Since the propagation of the traveling waves covers highly distributed brain regions, researchers have attempted to relate their functional significance to various aspects of the traveling waves. In particular, the directionality of traveling waves is believed to be functionally relevant (Klimesch et al., 2007; Fellinger et al., 2012; Patten et al., 2012; Bahramisharif et al., 2013). For example, Halgren and colleagues showed that, during wakefulness with open or closed eyes, oscillations recorded with intracortical electrodes from epilepsy patients propagated from antero-superior cortex toward postero-inferior occipital poles (Halgren et al.,

2019). However, in another intracortical study (Zhang et al., 2018), when subjects were instructed to complete a visual memory task, traveling waves in the θ-α band (2–15 Hz) propagated from posterior to anterior brain areas. This apparent forward direction of traveling waves was also reported in studies of so-called 'perceptual echoes', which constitute a direct index of sensory processing (VanRullen and Macdonald, 2012). Participants were stimulated with random (white-noise) luminance sequences, and the resulting impulse response function showed a long-lasting 10-Hz oscillation (or perceptual echo); importantly, the spatial distribution of the echo phase was organized as a traveling wave propagating from posterior to frontal sensors (Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019). It thus seems that the directionality of traveling waves could be task-dependent.

1.5.4 Predictive coding and oscillations

If the brain adopts a predictive coding strategy to operate, could brain oscillations serve to carry the corresponding predictions and error signals? In fact, many researchers have demonstrated such possibilities (Friston, 2019; Arnal and Giraud, 2012; Alamia and VanRullen, 2019; Engel et al., 2001).

One possible scheme is that prediction errors are transmitted by faster gamma oscillations, while predictions are conveyed by lower frequency bands like alpha or beta. This idea is formed mainly due to the empirical facts that higher frequencies such as gamma and beta are thought to carry the feedforward information while lower oscillations such as theta and alpha represent feedback signals (Bastos et al., 2015b; Michalareas et al., 2016; Van Kerkoerle et al., 2014). Thus it might be natural to link fast frequencies with predictive error signals and slow frequencies with prediction signals from a higher level. Functionally, the spectral asymmetries between predictions and error signals may be reasonable since they carry different types of signals as predictions are modulatory and errors signals are driven. Anatomically, earlier studies have shown that forward and backward projections can be differentiated based on their laminar patterns. Thus, it might be reasonable that those two channels transmit

two kinds of information.

Another scheme considers the spatial domain of oscillation, i.e., the traveling waves. Alamia and VanRullen (2019) suggest that the propagation of alpha oscillations might serve as the neural mechanism of predictive coding. They first built a two-layer model which implements predictive coding dynamics. The results show that the model could generate alpha-band oscillations with biologically plausible time constant and time delay. In their second experiment, they enlarged the model into multi-layers. As expected, the enlarged model could generate alpha oscillation propagating through the model layers. Remarkably, each layer of the enlarged model could correspond to large brain regions from occipital to frontal areas, thus the model could be compared with empirical data obtained from these areas. Both computational and experimental results showed that feedforward input results in traveling alpha waves from occipital to frontal regions, while feedback signals cause alpha waves passing in the opposite direction. That is, on a large scale, the prediction signals could be conveyed by downward alpha traveling waves; while feedforward prediction errors transmitted by upward alpha waves. Remarkably, the natural emergence of alpha oscillations suggests its high relevance with the predictive coding in the brain.

In summary, brain oscillations can serve as a potential candidate for the neural correlates of predictive coding implementation in the biological brain. However, it is still unclear which frequency band is involved, and how these oscillations could be related to the predictive process. Therefore, the current thesis will focus on this question and try to find the underlying mechanism.

1.6 Summary and objectives of the thesis

As the title suggested, the current thesis takes two aspects–in physical brains and deep neural networks–to evaluate the possibility of predictive coding as a unifying theory of brain function. From the perspective of brains, the predictive processing in brains as postulated by the theory can explain a wide range of neurophysiological and psychological phenomena as reviewed in section 1.3.2. If the dynamics in the physical brains are driven by a predictive coding strategy, then a natural question arises. What is the underlying neural mechanism? Or, which neural activities or factors can undertake the task of information transmission in terms of ascending prediction errors and descending prediction signals in the cortical hierarchy? In section 1.5, I discussed whether cortical oscillations including traveling waves could act as the underlying mechanism of predictive coding since they could play roles in a wide range of brain functions as reviewed in sections 1.5.2 and 1.5.3. Some researchers hold the idea that fast gamma oscillations could convey predictive errors in the cortical hierarchy; while much slower alpha oscillation could deliver descending prediction signals. Different opinions exist in Alamia and VanRullen (2019) showing both signals could be transmitted by alpha oscillation traveling between cortical areas. Thus it is still unclear whether cortical oscillations could underlie predictive coding and which frequency bands are involved.

Another aspect involves examining the theory in deep neural networks. Compared to empirical studies, the computational approaches can only provide indirect evidence. However, it is thought that the combination of both methods can provide revealing insights. The biological facts inspire the construction of artificial neural networks; in turn, the resulting performance of neural networks promotes more understanding of the brain dynamics. To take advantage of computational models powered by deep learning techniques (see section 1.4.1), the core idea is to construct a computational model with predictive coding dynamics and expect such network may display similar performance as the human brain if they share the same dynamic system, i.e. the predictive coding. Before being transformed into a model, the theory needs to be formalized properly. Section 1.3.3 displayed several possible ways of formulation in terms of the same theory–predictive coding. The construction and implementation process

can also vary. Section 1.4.2 reviewed a few different deep neural networks driven by predictive coding dynamics. As can be seen, multiple factors including the way of formulating a theory and the way to build and train a model may matter a lot when judging whether the theory, despite its original idea, could provide accurate predictions.

Our final goal is to answer whether predictive coding could act as a theory of brain function, which can be approached by finding its underlying neural mechanisms in the brain. To further concretize the problem, we can ask whether the oscillatory traveling waves could serve as the potential mechanism as suggested by section 1.5.4. However, even if we can show the traveling waves in a certain frequency band that may show similar activation patterns as suggested by predictive coding theory, this can only indicate a correlation relationship instead of a causal one. Thus, this possible result needs to be corroborated somewhere else. The method of the computational model or deep neural network can prove this in an opposite direction. That is, we can build a deep neural network with predictive coding dynamics. If such a network can show oscillations or traveling waves with biologically plausible parameters, we can be more confident about our hypothesis. We may even go further by showing both biological and artificial oscillations or traveling waves can affect a certain visual perception, say visual illusion, in the same way, since as a vision model, the deep neural network can deal with multiple visual tasks including visual recognition. However, before going that far, we may first need to prove the neural network with predictive coding can indeed show human-like performance given there are multiple ways to formulate, construct and train a model.

To summarize, our work can be divided into three specific studies: (i) to find the relevant neural mechanisms of predictive coding, we ask whether oscillatory traveling waves can serve as a possible relevant neural event by checking their processing characteristics in the brain; (ii) To obtain a human-like predictive coding network, we ask whether such a network could show human-like visual illusion; (iii) To further prove the underlying role of oscillations, we take the same predictive coding network and ask whether it can generate biologically plausible oscillations or even traveling waves?

2 Alpha traveling waves as potential neural correlates of predictive coding?

2.1 Chapter Introduction

The first concern in the thesis is about the neural implementation of predictive coding in biological brains. As we already know, in the hypothesized hierarchical model engaging in the predictive coding process, higher level sends predictions to a lower level; while error signals are conveyed in the opposite direction. Therefore our concrete question is to figure out which neural activities in the brain can carry both bidirectional signals?

Brain oscillations may serve as a possible candidate. The speculation may originate from such a fact that both oscillations and traveling waves have a strong explanation power towards a wide range of neurophysiological and psychophysical observations as reviewed in sections 1.3.2 and 1.5.2. More importantly, a more substantial relationship between predictive coding and oscillations may exist when considering the directionality of their message passing. It has been shown faster frequencies (e.g. gamma) could convey a forward message, while lower frequencies (e.g., alpha and beta) are related to the backward message passing (Bastos et al., 2015a,b; Arnal and Giraud, 2012). These spectral asymmetries in brain oscillation may shed light on the message passing in a predictive coding framework with faster frequency conveying ascending prediction errors and lower frequencies carrying predictions (Friston, 2019).

Alternatively, traveling waves may even serve a better role for operating the dynamics in predictive coding due to their natural directionality as they propagate between brain regions. This idea was proved by a modeling study (Alamia and VanRullen, 2019) where alpha oscillations traveled downwards the hierarchical model when only priors (predictions) were offered; while an opposite direction of alpha oscillations manifested with the sole presentation of visual input (error signals). Remarkably, under the predictive coding architecture, alpha oscillations emerge naturally with biologically plausible time constants and delays of neurons.

This indicates a close link between alpha traveling waves and predictive coding realization within the biological brain.

The first study was designed to provide empirical evidence for the modeling study by Alamia and VanRullen (2019) to finally prove the role of traveling waves in predictive coding. The general idea is to create two distinct processing conditions with each supporting traveling waves with a specific direction. In the modeling study, either pure predictions or error signals can be passed through the model and the resulting traveling waves can be observed. However, things might be complicated in the brain as the brain constantly generates predictions. It might be possible to only have predictions in the brain, for example, cutting off the sensory input; however, we cannot obtain a situation where only feedforward inputs are involved due to the predictions existing all the time. Therefore, in the first study, we employed a wave quantification method to sort forward and backward traveling waves and analyzed how they would be related to the messaging passing in a predictive coding framework.

2.2 Article 1

Turning the Stimulus On and Off Changes the Direction of Alpha Traveling Waves

Zhaoyang Pang, Andrea Alamia, and Rufin VanRullen. (2020) "Turning the Stimulus On and Off Changes the Direction of Traveling Waves." eNeuro, 7(6). https://www.eneuro.org/content/7/6/ENEURO.0218-20.2020

eNeuro

Research Article: Confirmation

Cognition and Behavior

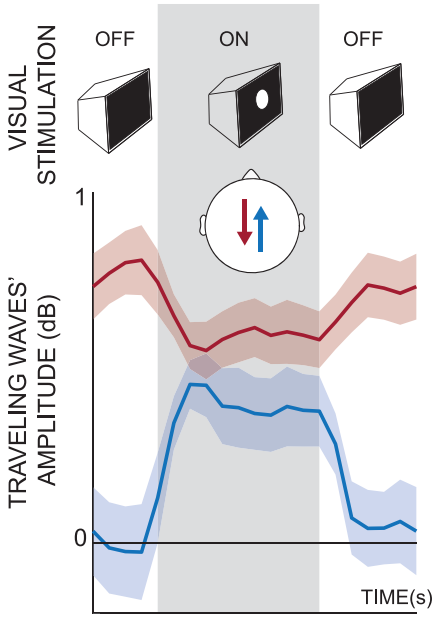
Turning the Stimulus On and Off Changes the Direction of α Traveling Waves

Zhaoyang Pang (庞兆阳),[1] Andrea Alamia,[1] and Rufin VanRullen[1,2]

https://doi.org/10.1523/ENEURO.0218-20.2020

[1]Centre de Recherche Cerveau et Cognition (CerCo), Centre National de la Recherche Scientifique, Université de Toulouse, Toulouse 31052, France and [2]Artificial and Natural Intelligence Toulouse Institute (ANITI), Toulouse 31000, France

Visual Abstract



Traveling waves have been studied to characterize the complex spatiotemporal dynamics of the brain. Several studies have suggested that the propagation direction of α traveling waves can be task dependent. For example, a recent electroencephalography (EEG) study from our group found that forward waves (i.e., occipital to frontal, FW waves) were observed during visual processing, whereas backward waves (i.e., frontal to occipital,

Significance Statement

Several electroencephalography (EEG) studies have suggested that the propagation direction of α traveling waves can be task dependent; however, these recordings were obtained from different experimental sessions and different groups of subjects. Here, we conducted a human EEG experiment with both visual processing and resting state combined into each single trial. Forward waves (FW waves) from occipital to frontal regions, absent during rest, emerged as a result of visual processing, while backward waves (BW waves) dominated in the absence of visual inputs. Importantly, during visual processing, both FW and BW α waves were present and modulated by stimulation type (static or dynamic), but they were negatively correlated over time.

BW waves) mostly occurred in the absence of sensory input. These EEG recordings, however, were obtained from different experimental sessions and different groups of subjects. To further examine how the waves' direction changes between task conditions, 13 human participants were tested on a target detection task while EEG signals were recorded simultaneously. We alternated visual stimulation (5-s display of visual luminance sequences) and resting state (5 s of black screen) within each single trial, allowing us to monitor the moment-to-moment progression of traveling waves. As expected, the direction of α waves was closely linked with task conditions. First, FW waves from occipital to frontal regions, absent during rest, emerged as a result of visual processing, while BW waves in the opposite direction dominated in the absence of visual inputs, and were reduced (but not eliminated) by external visual inputs. Second, during visual stimulation (but not rest), both waves coexisted on average, but were negatively correlated. In summary, we conclude that the functional role of α traveling waves is closely related with their propagating direction, with stimulus-evoked FW waves supporting visual processing and spontaneous BW waves involved more in top-down control.

Key words: α oscillations; predictive coding; traveling waves; visual processing; waves propagating direction

Introduction

Neural oscillations at various temporal frequencies are ubiquitous in the human brain, and in the spatial domain, an increasing number of studies suggest that these oscillations could be organized as traveling waves across brain regions. The existence of traveling waves has been reported across multiple species (Ermentrout and Kleinfeld, 2001; Sato et al., 2012), at differing scales of measurements (Muller et al., 2018), and under various stimulation conditions (Nauhaus et al., 2012; Sato et al., 2012). Since the propagation of the traveling waves covers highly distributed brain regions, researchers have attempted to relate their functional significance to various aspects of the traveling waves. In particular, the directionality of traveling waves is believed to be functionally relevant (Klimesch et al., 2007a; Fellinger et al., 2012; Patten et al., 2012; Bahramisharif et al., 2013). For example, Halgren and colleagues showed that, during wakefulness with open or closed eyes, α oscillations recorded with intracortical electrodes from epilepsy patients propagated from antero-superior cortex toward postero-inferior occipital poles (Halgren et al., 2019). However, in another intracortical study (Zhang et al., 2018), when subjects were instructed to complete a visual memory task, traveling waves in the θ-α band (2–15 Hz) propagated from posterior to anterior brain areas. This apparent forward direction of traveling waves was also reported in studies of so-called "perceptual echoes," which constitute a direct index of sensory processing (VanRullen and Macdonald, 2012). Participants were stimulated with random (white-noise) luminance sequences, and the resulting impulse response function showed a long-lasting 10-Hz oscillation (or perceptual echo); importantly, the spatial distribution of echo phase was organized as a traveling wave propagating from posterior to frontal sensors (Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019). It thus seems that the directionality of traveling waves could be task dependent. To clarify the traveling direction with respect to various experimental conditions, a recent study (Alamia and VanRullen, 2019) from our group simulated α oscillations as a cortical traveling wave within a predictive coding framework. The predictive coding framework characterizes a hierarchical network where higher levels of brain regions predict the activity of lower levels, and the unexplained residuals (i.e., prediction errors) are passed back to higher layers. The study revealed that the recursive nature of predictive coding not only gave rise to α oscillations but also explained their propagating dynamics. Remarkably, when feeding with visual inputs (e.g., white noise), simulated α oscillations propagated from lower level to higher level, while simulating resting state gave rise to feedback waves.

The computational study suggests that the directionality of traveling waves could be closely linked with task conditions (visual processing vs rest state) and is supported by human electroencephalography (EEG) studies where participants were instructed to monitor a visual luminance sequence or keep their eyes closed. However, those human experiments were conducted separately within different experimental sessions and different groups of participants, and it is thus difficult to infer a direct relationship between the task condition and waves' direction. To verify the predictions of the computational work and to systematically examine how the waves' direction changes from one task condition to another, the current EEG study was designed to incorporate stimulus-on periods (visual processing) and stimulus-off periods (resting state) within each single trial, by which we could trace the moment-to-moment changes of the waves' direction caused by task conditions in a consistent way.

Received May 21, 2020; accepted October 23, 2020; First published November 6, 2020.

The authors declare no competing financial interests.

Author contributions: R.V. designed research; Z.P. and A.A. performed research; Z.P. and A.A. analyzed data; Z.P. wrote the paper.

This work was supported by the European Research Council Consolidator Grant P-CYCLES 614244, the ANR (Agence Nationale de la Recherche) OSCI-DEEP Grant ANR-19-NEUC-0004 and an ANITI Chair Grant ANR-19-PI3A-0004 to R.V. Z.P. is supported by the China Scholarship Council Grant 201806620059.

Correspondence should be addressed to Rufin VanRullen at rufin.vanrullen@cnrs.fr.

https://doi.org/10.1523/ENEURO.0218-20.2020

Copyright © 2020 Pang et al.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

Materials and Methods

Participants

A total of 14 subjects participated in this experiment. One subject was rejected because of a technical problem

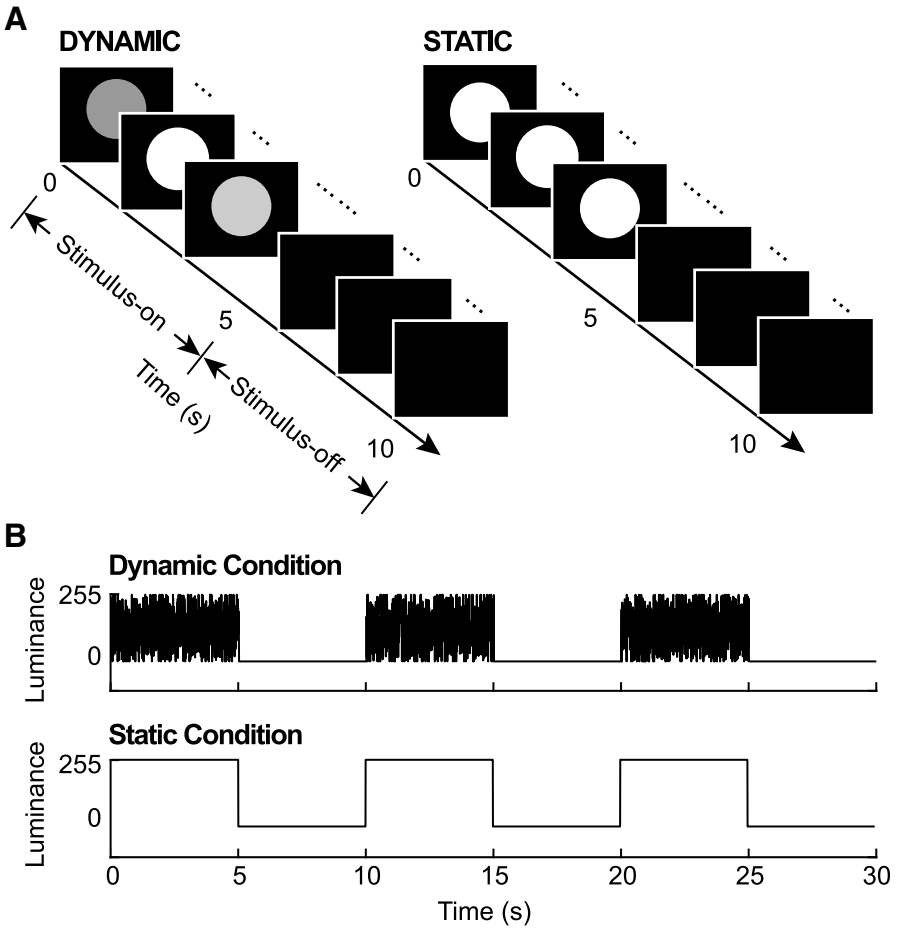


Figure 1. Experiment design. ***A***, Two types of trials were included in this experiment. For static trials, the luminance of visual input was held constant at a value of 255 (full contrast), while for dynamic ones, the luminance changed randomly from 0 to 255 on each screen refresh. In both cases, luminance sequences were displayed for the first 5 s (stimulus-on period), then followed by 5 s of blank screen (stimulus-off period). ***B***, Schematic diagrams of two subblocks for both dynamic and static conditions. Each subblock contained three identical trials, which made up a 30-s-long time course.

during the experimental recording, leaving 13 subjects (six females; mean age 25.57, range 21–31; two left-handed) for inclusion in the analysis. All participants reported no history of epileptic seizures or photosensitivity and they had normal or corrected to normal vision. Before starting the experiment, all participants gave written informed consent as specified by the Declaration of Helsinki. The study was performed under the guidelines for research according to author's research institute at the Centre de Recherche Cerveau et Cognition and the protocol was approved by the committee Comité de protection des Personnes Sud Méditerranée 1 (ethics approval number N° 2016- A01937-44).

Stimuli generation

Visual stimuli were generated using MATLAB scripts and presented using the Psychophysics Toolbox (Brainard, 1997). The stimuli were displayed on a cathode ray monitor in a dark room, positioned 57 cm from the subjects, with a refresh rate of 160 Hz and a resolution of 800 × 600 pixels. We used two types of visual luminance sequences (Fig. 1) as visual inputs: dynamic (or white-noise) and static stimulation. For the white-noise sequences, the power spectrum was normalized to have equal power at all frequencies (up to 80 Hz). The resulting luminance of white-noise sequences ranged from black (0.1 cd/m^2) to white (59 cd/m^2), whereas the static ones were held constant with full contrast (59 cd/m^2). Luminance sequences were displayed for 5 s within a disk of 3.5° radius which was centered at 7.5° above a center white dot on a black background.

Experimental design

Subjects were instructed to perform a visual detection task. During the experiment, three identical trials (either static or dynamic) were displayed in a row, grouped into a subblock (Fig. 1*B*). Before each subblock, a green center dot was displayed until subjects pressed the space bar to indicate their readiness. The green dot then disappeared and was followed by those three trials after a time interval of 200–300 ms. A prototypical trial started (Fig. 1*A*) with 5 s of luminance sequences (either dynamic or static) in a disk above a white fixation dot at the center of the screen and then 5 s of blank screen. That is, each trial contained

a stimulus-on period and a stimulus-off period, which allowed us to investigate the moment-to-moment changes of traveling waves when shifting from one task condition to another. Observers were asked to keep their fixation throughout the trial. Also, during visual stimulation (stimulus-on period), observers needed to covertly attend the disk to detect a brief square target (decreased luminance) inside the disk.

Two types of trials lead to two corresponding subblocks, dynamic or static, which were presented alternatively and also counterbalanced within subjects. Targets (1 s) appeared at a random time (uniform distribution) from 0.25 s after the onset to 0.25 s before the offset of luminance sequences on a random 20% of trials. The square target luminance was adjusted according to each subject by a staircase procedure using the Quest function (Watson and Pelli, 1983) to ensure 80% detection rate. In dynamic trials, the luminance of the square target fluctuated according to the white-noise sequences, but with a lower contrast compared with the rest of the disk. The experiment was composed of five sessions of 10 experimental blocks of six trials (i.e., two subblocks) each, with a total duration of ~1 h.

EEG recording and preprocessing

Continuous brain activity was recorded from the subjects using a 64-channel active BioSemi EEG system, with 1024-Hz digitizing sample rate and three additional ocular electrodes. Custom scripts in the EEGlab toolbox (Delorme and Makeig, 2004) were applied to the pre-processing steps, during which both target-present and target-absent trials were included. We first rejected the noisy channels and then the data were offline downsampled to 160 Hz. In order to remove power line artefacts, a notch filter (47–53 Hz) was applied. We applied an average-referencing and removed slow drifts by applying a high-pass filter (>1 Hz). Data epochs were created around −0.5 to 10 s around the trial onset, and EEG activity was corrected by subtracting the baseline activity from −0.5 to 0 s before trial onset. Finally, the data were screened manually for eye movements, blinks and muscular artefacts and whole epochs were rejected as needed.

Wave quantification

In order to quantify the presence of traveling waves in EEG signals and assess the propagation direction, we adopted a wave quantification method from our precious studies (Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019), which is described in Figure 2. For each subject, every trial (10 s long with 0.5-s baseline) was divided into 20 time bins by a sliding window of 1 s (with 500-ms overlap). For each time bin, we stacked EEG signals from seven midline electrodes (from posterior to frontal: Oz, POz, Pz, CPz, Cz, FCz, Fz) to form a 2D (electrode-time) map. To computationally quantify the waves' amount, we used a 2D-FFT (2-D fast Fourier transform) transform for each 2D map. This transform results in temporal frequencies along the horizontal axis as well as spatial frequencies along the vertical axis. The horizontal midline indicates stationary oscillations with no spatial propagation, while the upper and bottom quadrants reflect forward-propagating (FW) and backward-propagating (BW) waves, respectively. We extracted the max value within the α band temporal frequencies (8–13 Hz) from the upper quadrant of the 2D-FFT as the FW value for this time window, and the max value (also within the α band) from the lower quadrant as the BW value. After repeating this procedure over all 20 time bins, we finally obtained two curves representing the dynamic changes of FW and BW waves along time.

To assess statistical significance of traveling waves, we used a non-parametric test. Specifically, we shuffled the electrodes' order 100 times for each time bin, thereby eliminating any spatial organization of the oscillatory signals (including traveling waves). For this surrogate data set, we repeated the same 2D-FFT procedure as described above. Since the shuffling procedure only eliminated the spatial structure but left intact the oscillatory power of EEG signals, the resulting FW and BW curves (Fig. 2*C*), based on the maximum power in each quadrant, could still fluctuate across time: oscillatory power was relatively suppressed during stimulus-on periods, then increased in the absence of visual input. These power fluctuations in the surrogate data, however, were similar in the FW and BW directions (as expected because of the shuffling procedure). In order to focus on the differences between real and surrogate data, we corrected the real wave patterns by dividing their values by the corresponding surrogate patterns, and expressing the result in dB units (Fig. 2*D*).

Analysis

We first conducted one-sample *t* tests against zero for both BW and FW waves separately to confirm their presence at each time point [corrected for multiple comparisons via false discovery rate (FDR), $\alpha = 0.05$]. Second, we examined differences between the waves across the different experimental conditions. For this, we conducted a within-subject three-factor repeated measure ANOVA: CONDITIONS (static vs dynamic visual stimulation) × WAVES (FW vs BW) × TIME BINS (20). To clearly examine the influence of tasks (visual processing vs rest state), we also grouped all time points within the stimulus-on and stimulus-off periods and conducted another ANOVA with factors CONDITIONS (static vs dynamic) × WAVES (FW vs BW) × TASKS (stimulus-on vs stimulus-off).

Results

Stimulus-evoked FW waves and spontaneous ongoing BW waves

Figure 3*A* illustrates the evolution of FW (blue) and BW (red) waves as a function of time under dynamic and static visual stimulation conditions, averaged across all subjects. The waves' traces in both plots show similar patterns overall: BW waves are relatively high during both stimulus-on and stimulus-off periods (with a decrease during stimulation), while FW waves seem to only emerge

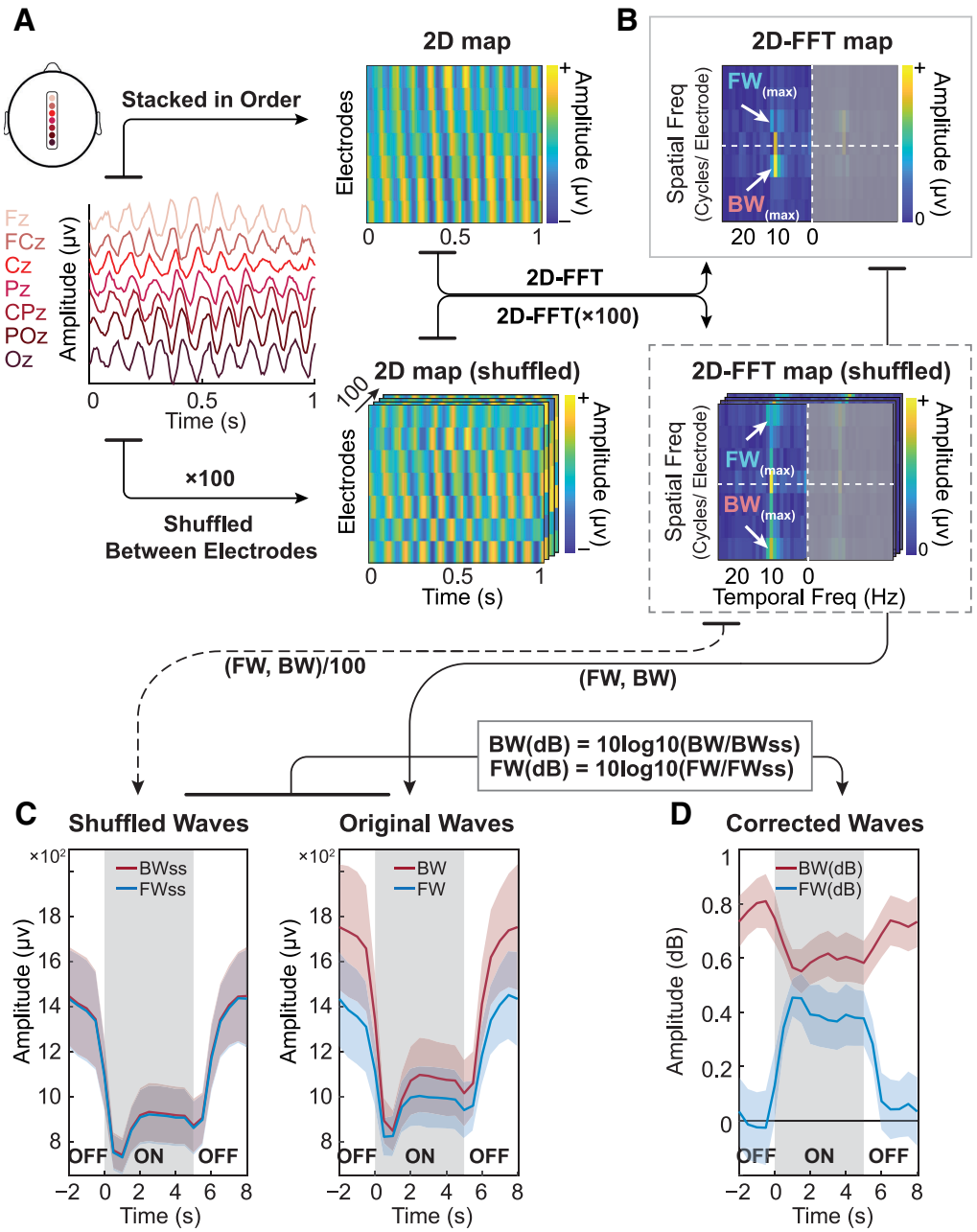


Figure 2. α Band traveling waves in raw EEG signals. ***A***, left, Seven midline electrodes of the 10–20 system are ordered from posterior to anterior (Oz to Fz) and backward traveling waves (BW) can be observed in the 1-s-long time window. Right, A 2D map of the same data with electrodes stacked in order and with amplitude color coded (top). To statistically quantify the waves' direction, we employed a non-parametric test by shuffling the electrodes' order for each time window 100 times. The resultant surrogate 2D maps eliminate the spatial structure of the original signals, including their original propagating direction (bottom). ***B***, Temporal frequencies (*x*-axis) and spatial frequencies (*y*-axis) for both real and surrogate data are obtained by computing a 2D-FFT. The temporal frequencies were computed up to 80 Hz, but only displayed until 25 Hz for illustration purposes. Since the 2D-FFT gives symmetrical results around the origin, we only focused on the left part of the plot. The maximum value in the upper quadrant represents the strength of forward traveling waves, while the maximum value in the lower quadrant quantifies the strength of feedback traveling waves. ***C***, left, For surrogate data, we averaged the 100 surrogate values separately for BW and FW signals and for each time bin. Colored shaded area stands for SEM across subjects; the stimulus-on period was shifted to the center part (gray shaded area) for better visualization of these dynamics around stimulus onset and offset. Right, Similar time courses were obtained for the real data. ***D***, The surrogate line plots were used as a baseline, mostly reflecting the background (α) oscillatory power. After correcting for these baseline fluctuations (and expressing the result in dB, as per the equations), we obtained a measure of the dynamics of FW and BW waves. Colored shaded area stands for SEM across subjects.

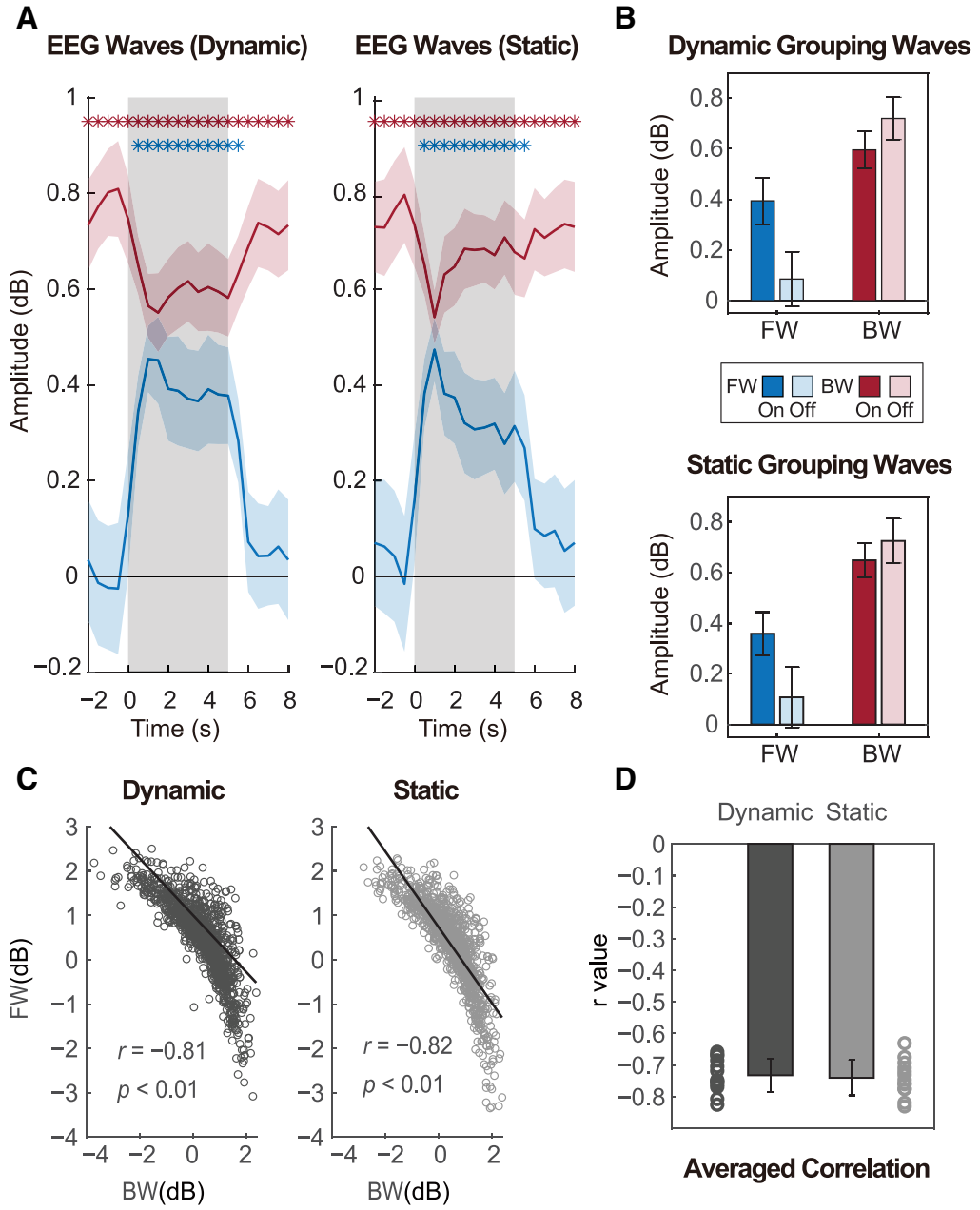


Figure 3. Amplitude of FW and BW waves and their correlation. ***A***, The two plots show how waves evolve over time under dynamic (white-noise visual sequences) and static (full contrast visual sequences) task conditions. Blue and red asterisks represent separately the significant time points (corrected for multiple comparisons by FDR) for FW waves (blue) and BW waves (red) when compared with zero. ***B***, To better compare wave patterns, waves were grouped over time points within each period type (stimulus ON/OFF). ***C***, For stimulus-on periods, we computed the correlation between BW and FW values under dynamic stimulation (left) and static stimulation (right). The two scatter plots show results from a representative subject. Each dot represents a pair of FW and BW values for each single time bin (1 s) and black lines are regression lines. ***D***, Bar graphs show averaged correlation (r values) across all subjects for dynamic and static conditions. The corresponding circles are individual results.

after the onset of visual stimulation, and disappear after the offset of visual inputs. In other words, the occurrence of FW waves is highly dependent on external stimulation while BW waves exist both in the presence and absence of stimulation. Therefore, we propose that FW waves are associated with visual processing (e.g., as a stimulus-evoked wave), while BW waves reflect ongoing spontaneous or endogenous activity. To support this, we conducted one-sample *t* tests against zero ($p < 0.05$, corrected for multiple comparisons by FDR) for the two waves separately at each time point. BW waves were significant during the entire time course (significant values are marked with asterisks in Fig. 3*A*). However, FW waves were only significant from 0.5 to 5.5 s in both dynamic and static conditions.

While the directionality of the waves along the occipito-frontal midline appears clear, one might wonder whether and how waves propagate over the rest of the scalp. Figure 4 illustrates the directional propagation along eight

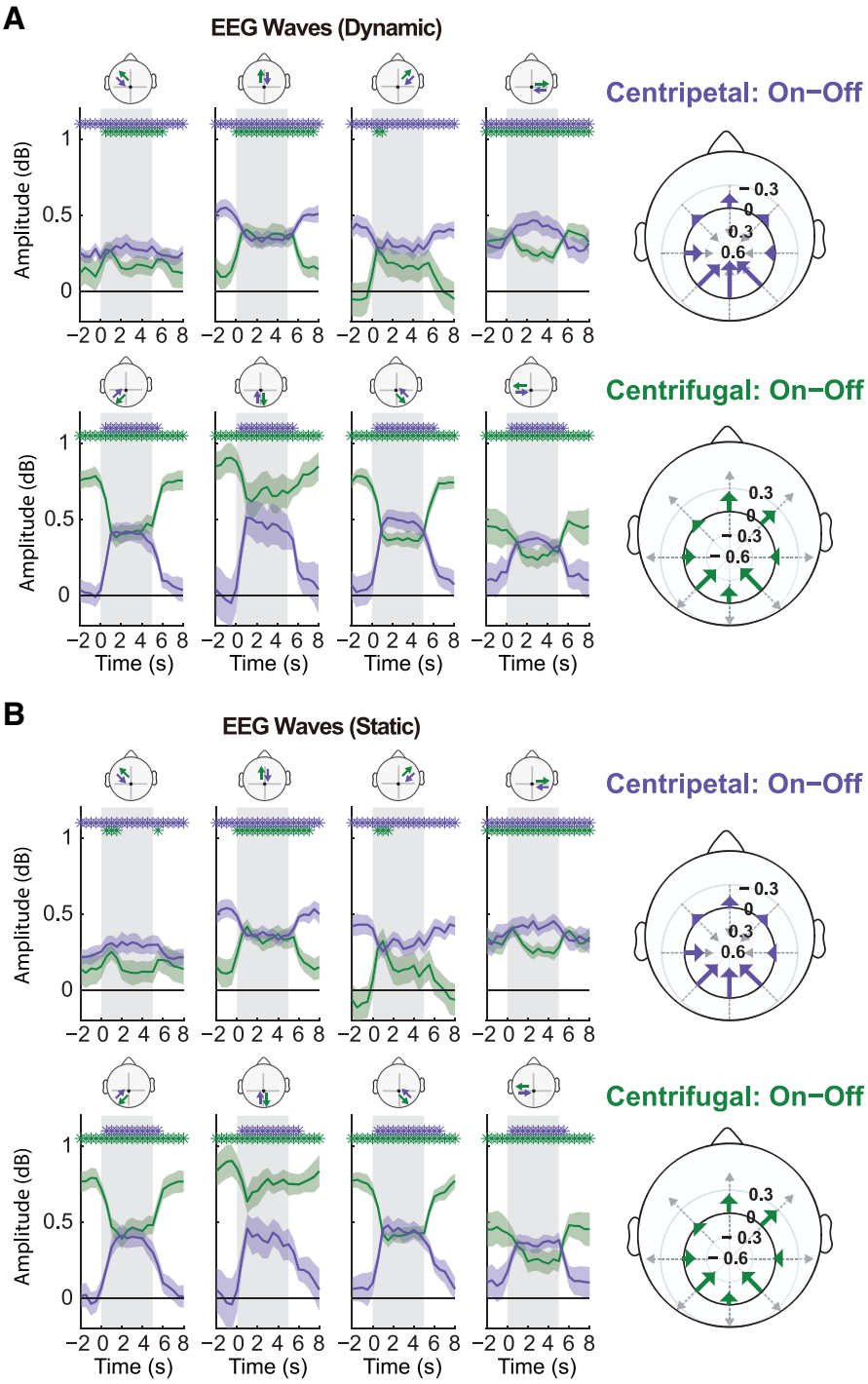


Figure 4. Traveling waves across the scalp. ***A***, left, Changes of wave amplitude over time under dynamic stimulation conditions. Around the central electrode CPz, we selected eight lines of electrodes to cover the whole scalp. As before, we computed two waves in opposite directions for each path (see inset topography above each subplot); here they are sorted as centripetal (toward CPz, in purple) and centrifugal (away from CPz, in green) waves. Each path counted only five electrodes, which were interpolated into seven sampling points, so as to remain comparable with our main analysis (see Fig. 3*A*). Right, Polar representation of the difference of wave amplitude between stimulus-on (gray shaded regions in left subplots) and stimulus-off period (white regions). Each wave is represented as an arrow, centered and aligned on the corresponding path, with a length proportional to its amplitude difference (with negative differences pointing in the opposite direction). The top image represents the vector field for centripetal waves, the bottom one for centrifugal waves. In both cases, the net effect of visual stimulation is to increase the forward propagation of the waves (and/or decrease their backward propagation, with little effect along the lateral axes). ***B***, Traveling waves under static stimulation conditions (notations and conclusions as in ***A***).

distinct paths on the scalp. Specifically, eight lines of five electrodes each were selected around the central electrode CPz. Since we could derive two opposite waves along each line, and not all of them included a clear FW versus BW direction (i.e., lateral lines), we sorted the waves in two groups: centripetal waves, toward CPz (purple), and centrifugal waves, away from CPz (green). Both the dynamic and static conditions showed very similar patterns: all the waves running in the occipito-frontal direction, including vertical but also diagonal lines, showed relatively large and task-dependent fluctuations, going from near-zero amplitude during rest to strong positive values during stimulation. The opposite (fronto-occipital) direction in each line tended to show significant waves throughout each trial, but stronger during rest and decreasing during visual stimulation. Finally, lateral directions of propagation displayed the smallest amount of fluctuations, regardless of the (centripetal or centrifugal) direction. This overall pattern was further confirmed by the polar plots on the right of the figure, obtained by subtracting averaged waves during stimulus-off periods from the corresponding waves during stimulus-on periods. In all cases, the changes in wave amplitude caused by the visual stimulation onset mainly lie along the occipital-frontal direction. Thus, our initial result with midline electrodes appears to be representative of the behavior of traveling waves across the entire scalp.

Both FW and BW waves are task dependent

After establishing the presence of FW waves during visual stimulation, and BW waves during both visual stimulation and resting state, we further examined the properties of both waves under various task conditions, using a three-way repeated measures ANOVA with factors CONDITIONS (dynamic/static), WAVES (FW/BW) and TIME BINS (20 values). This revealed main effects for TIME BINS ($F_{(19,228)} = 19.083$, $p < 0.001$, $\eta_p^2 = 0.614$) and WAVES ($F_{(1,12)} = 7.048$, $p = 0.021$, $\eta_p^2 = 0.37$), a significant two-way interaction for WAVES $\times$ TIME BINS ($F_{(19,228)} = 9.002$, $p < 0.001$, $\eta_p^2 = 0.429$), as well as a significant three-way interaction ($F_{(19,228)} = 3.103$, $p < 0.001$, $\eta_p^2 = 0.205$).

The CONDITIONS $\times$ TIME BINS interaction reached significance for FW waves ($F_{(19,228)} = 2.698$, $p < 0.001$, $\eta_p^2 = 0.184$) at time points 2, 3.5, and 5 s, and for BW waves ($F_{(19,228)} = 2.928$, $p < 0.001$, $\eta_p^2 = 0.196$) at time points 2–4 and 5 s. That is, the time course of FW waves showed less power for static stimulation at certain time points. BW waves were also influenced by the stimulation type but with increasing power for static stimulation toward the later part of each stimulation period (Fig. 3*A*). We speculate that both waves may be influenced by stimulus complexity since static stimuli are much simpler and more predictable compared with dynamic white-noise luminance sequences.

FW waves were only present during visual processing, while BW waves existed during both task conditions (visual processing vs rest). To further examine whether BW waves showed significant differences associated with the tasks, another three-way repeated measures ANOVA was conducted with factors CONDITIONS (dynamic/static), WAVES (FW/BW) and TASKS (stimulus-on/off). That is, the TIME BINS factor (20 values) was replaced with the TASKS factor (two values). The wave amplitudes were averaged over time bins (separately for the stimulus-on and stimulus-off conditions). This time, we did not obtain a significant three-way interaction (Fig. 3*B*). Instead, the ANOVA revealed a significant two-way interaction for WAVES $\times$ TASKS ($F_{(1,12)} = 18.056$, $p = 0.001$, $\eta_p^2 = 0.6$). As expected, the main effect of CONDITIONS was significant ($F_{(1,12)} = 25.95$, $p < 0.001$, $\eta_p^2 = 0.684$) for FW waves, similar to the result of one-sample *t* tests above (Fig. 3*A*). For BW waves, the main effect of CONDITIONS was also significant ($F_{(1,12)} = 7.196$, $p = 0.02$, $\eta_p^2 = 0.375$) with higher BW power in the absence of visual inputs. The modulation of BW waves by visual stimulation is in line with other studies showing that spontaneous traveling waves could be suppressed by external inputs (Patten et al., 2012; Sato et al., 2012).

In summary, FW waves appear to be caused by external visual stimulation, while BW waves can originate spontaneously but could be reduced (yet not eliminated) by the presence of visual stimulation. During visual stimulation, both waves are present and modulated by the type of visual inputs, with lower FW but higher BW waves' power for simpler (static) sensory stimulation.

FW and BW waves are negatively related during visual stimulation

During visual stimulation (stimulus-on periods), both FW and BW waves appear to be simultaneously present (Fig. 3*A*). However, the average traveling wave's behavior does not necessarily reflect the instantaneous state of the brain and its dynamics: FW and BW waves may by truly equally present at each moment in time, or they may tend to happen in alternation, at distinct moments in time. To further examine the relationship between them, we assessed the moment-to-moment correlation between FW and BW waves for each stimulus condition. Figure 3*C* shows scatter plots from a representative subject. For each condition, each 1-s time bin window produces one pair of FW and BW wave values, i.e., a single dot in the plot. To discard the common influence of oscillatory amplitude fluctuations on both FW and BW traveling waves, we correct each wave's value by its corresponding averaged surrogate value (obtained by shuffling the electrodes' order 100 times, as explained in Fig. 2*A*). The correlation between the resulting FW and BW values in dB units showed a clear and significant ($p < 0.01$) negative trend, for both the dynamic and static stimulus conditions. This means that when FW waves were stronger, BW waves tended to be weaker, and vice-versa. Figure 3*D* gives the average correlation across all subjects: significant negative correlation between FW and BW waves can be observed for both the dynamic (mean = -0.732 ± 0.053, $t_{(12)} = -49.371$, $p < 0.001$, 95% confidence interval (CI): -0.765 to -0.7) and static conditions (mean = -0.74 ± 0.056, $t_{(12)} = -47.644$, $p < 0.001$, 95% CI: -0.774 to -0.706).

Discussion

Based on EEG data from human participants, we demonstrated that the direction of α traveling waves (8–13 Hz)

is task dependent, confirming suggestions from prior studies (Zhang et al., 2018; Alamia and VanRullen, 2019; Halgren et al., 2019; Lozano-Soldevilla and VanRullen, 2019), and verifying the predictions of our own modeling study on the generation and propagation of α oscillations (Alamia and VanRullen, 2019). Specifically, we characterized FW waves traveling from occipital to parietal regions elicited by visual stimulation, and BW waves in the reversed direction dominating during rest state. Furthermore, the presence of external visual stimulation reduced BW waves (Fig. 3*A*), which is in line with other studies on spontaneous traveling waves (Patten et al., 2012; Sato et al., 2012). Lastly, during visual stimulation, FW waves and BW waves were present and modulated by stimulation type (static or dynamic), but they were negatively correlated over time.

Contributions of the current study

It should be emphasized that the current experimental design directly contrasted the conditions of visual processing and resting state within each trial. Previously, a number of studies had examined traveling waves under various single-task conditions, including visual stimulation (Nauhaus et al., 2012; Muller et al., 2014; Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019), sleep (Muller et al., 2016), or quiet wakefulness (Alamia and VanRullen, 2019; Halgren et al., 2019). While these experiments confirmed the existence of traveling waves, they did not make it possible to track how the waves change from one condition to another. Because of the within-subject design in the present study, we found that the waves' direction is highly sensitive to the task conditions.

Compared with previous studies using dynamic white noise sequences as visual stimulation (Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019), we also included a simpler type of visual stimulation: static luminance sequences. The results showed that although these two stimulus types evoked similar FW and BW waves, toward the later stages of visual stimulation, the BW wave power increased at time points 2–4 and 5 s and FW wave power decreased at 2, 3.5, and 5 s. This may be because of the relative simplicity of static inputs compared with the dynamic ones. The same factor, and the longer trial durations, may also explain why we measured less FW power overall compared with our prior studies (Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019). Simpler stimuli and longer trials may result in subjects being less engaged in the task, and consequently in weaker FW waves overall. Future studies should explore whether the overall amount of FW power can be increased by parametric manipulations of the task or of the experimental stimuli or screen background.

Unlike prior studies measuring a single traveling wave direction from phase gradients over certain brain regions (Zhang et al., 2018; Halgren et al., 2019), we here derived two opposite components of the waves' direction from the pattern of brain activity within each time window, and quantified their strength. Previous studies have shown that traveling waves can propagate in different directions (Alexander et al., 2009; Patten et al., 2012), and that the co-existence of two opposite waves may cause a loss of wave information (Alexander et al., 2013). For example, under cortical states like sleep or resting state, traveling waves have often been reported to propagate in a frontal-to-occipital direction (Massimini et al., 2004; Alamia and Vanrullen, 2019). However, traveling waves are less frequently observed under more complex cognitive states (Alexander et al., 2013), this may be caused by the interference of waves propagating in opposite directions, while their direction is characterized as a single value. Instead, our analysis method used in the current study independently quantifies waves propagating in the two directions. Also, the separation of FW and BW waves' components contributed to reveal their distinct functional roles. We revealed a closer link between FW waves and visual processing as an evoked wave, since FW waves emerged at the onset of visual input and disappeared right after the offset (Fig. 3*A*); meanwhile, BW waves were more related to the resting state, acting as a spontaneous wave.

An explanation under the predictive coding framework

The generation and directionality of traveling waves can tentatively be interpreted within the predictive coding framework (Rao and Ballard, 1999). In our previous work (Alamia and VanRullen, 2019), researchers built a seven-level hierarchical model of visual cortex with bidirectional connectivity implementing predictive coding. Within the hierarchy, higher levels predicted the activity of lower ones through inhibitory feedback, and lower levels sent the prediction error via feedforward excitation to the higher layers to correct their prediction. With biologically plausible parameters (neural time constants, communication delays), this model produced α rhythms traveling through the hierarchy. The waves could travel in the FW direction when the model was presented with visual inputs, and in the BW direction in the absence of inputs (while the model was processing "top-down priors" instead of bottom-up sensory signals).

In this context, it is reasonable to infer that FW waves carry "residual error" signals (the difference between the actual visual inputs and the prediction from higher-level regions), while BW waves carry the prediction signals. Remarkably, the current results that FW waves emerged only during visual stimulation and BW waves were dominant in the resting state agree with this framework. On the other hand, the negative correlation across time between FW waves and BW waves during visual stimulation may reflect the dynamics of predictive coding mechanism. That is, stronger prediction signals within BW waves are associated with weaker prediction errors carried by FW waves and vice versa. Moreover, in the static condition, BW waves increased but FW waves decreased significantly at the later stages of visual stimulation, indicating that prediction information becomes stronger while error signals weaken over time. This was not the case in the dynamic condition, which has much more complex (and unpredictable) stimulus temporal structure, leading to less precise prediction signals and larger error signals.

Spontaneous BW waves may reveal top-down control

Spontaneous ongoing waves have been reported in the cortex under anesthesia or quiet wakefulness (Petersen et al., 2003; Sakata and Harris, 2009; Alamia and VanRullen, 2019). The current study points to BW waves as spontaneous waves, given their existence under resting state. Besides, the significant reduction of BW waves because of the presence of visual inputs also agrees with other studies on spontaneous traveling waves (Patten et al., 2012; Sato et al., 2012). This reduction could be explained by the desynchronization caused by visual processing, since spontaneous activity measured during quiet wakefulness may reflect synchronized cortical states (Harris and Thiele, 2011). On the other hand, given the spatial extent of traveling waves across distributed cortical regions, their functional role may entail long-range information integration (Sato et al., 2012; Halgren et al., 2019). In particular, it is speculated that BW waves may participate in the organization of top-down or feedback information flow (Van Kerkoerle et al., 2014; Halgren et al., 2019). This is in line with the dominance of α band activity in the waves, a frequency which is typically associated with top-down control (Klimesch et al., 2007b; Jensen et al., 2012).

Stimulus-evoked FW waves are associated with bottom-up sensory processing

In the current study, we measured α FW waves which were directly linked with visual processing (and absent during rest). This direct link is also supported by our prior studies of perceptual echoes: since these echoes are measured by cross-correlation with the visual input sequence, they can be viewed as a direct reflection of visual processing (VanRullen and Macdonald, 2012; Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019). Recent work from our group found that these perceptual echoes propagate from occipital to parietal regions in a forward direction (Alamia and VanRullen, 2019; Lozano-Soldevilla and VanRullen, 2019). Although further research is needed to test whether FW waves also contribute to sensory processing in other modalities (like audition or touch), FW waves may serve to integrate the information flow along the bottom-up path. This is consistent with the "scanning hypothesis" proposed by Pitts and McCulloch (1947), suggesting that the α rhythm repeatedly scans the visual cortex. The bidirectionality of α traveling waves found in the current study may help to clarify an apparent contradiction between the conventionally postulated inhibitory role of α oscillations (Jensen and Mazaheri, 2010; Bonnefond and Jensen, 2012), and their reported implication in sensory processing (Varela et al., 1981; VanRullen, 2016). Inhibition may be carried by the BW component of α oscillations as mentioned above, whereas, the FW component may reflect the positive relation between α and sensory processing.

In summary, the current study corroborated the predictions from our prior EEG and modeling study (Alamia and VanRullen, 2019). It showed that FW and BW waves are inversely related to sensory processing, and may characterize opposite directions of information flow in the brain hierarchical system. Importantly, the transitions between FW and BW waves were observed within single trials and for the same human subjects. First, FW waves travel from occipital to frontal regions during visual processing, while BW waves are spontaneously generated and travel in the opposite direction, likely reflecting a feedback process. Second, during visual stimulation, both FW and BW waves exist on average, but are negatively correlated across time, suggesting that they reflect distinct functions that may draw on common brain resources.

References

Alamia A, VanRullen R (2019) Alpha oscillations and traveling waves: signatures of predictive coding? PLoS Biol 17:e3000487.

Alexander DM, Flynn GJ, Wong W, Whitford TJ, Harris AWF, Galletly CA, Silverstein SM (2009) Spatio-temporal EEG waves in first episode schizophrenia. Clin Neurophysiol 120:1667–1682.

Alexander DM, Jurica P, Trengove C, Nikolaev AR, Gepshtein S, Zvyagintsev M, Mathiak K, Schulze-Bonhage A, Ruescher J, Ball T, van Leeuwen C (2013) Traveling waves and trial averaging: the nature of single-trial and averaged brain responses in large-scale cortical signals. Neuroimage 73:95–112.

Bahramisharif A, van Gerven MAJ, Aarnoutse EJ, Mercier MR, Schwartz TH, Foxe JJ, Ramsey NF, Jensen O (2013) Propagating neocortical gamma bursts are coordinated by traveling alpha waves. J Neurosci 33:18849–18854.

Bonnefond M, Jensen O (2012) Alpha oscillations serve to protect working memory maintenance against anticipated distracters. Curr Biol 22:1969–1974.

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21.

Ermentrout GB, Kleinfeld D (2001) Traveling electrical waves in cortex. Neuron 29:33–44.

Fellinger R, Gruber W, Zauner A, Freunberger R, Klimesch W (2012) Evoked traveling alpha waves predict visual-semantic categorization-speed. Neuroimage 59:3379–3388.

Halgren M, Ulbert I, Bastuji H, Fabó D, Erőss L, Rey M, Devinsky O, Doyle WK, Mak-McCully R, Halgren E, Wittner L, Chauvel P, Heit G, Eskandar E, Mandell A, Cash SS (2019) The generation and propagation of the human alpha rhythm. Proc Natl Acad Sci USA 116:23772–23782.

Harris KD, Thiele A (2011) Cortical state and attention. Nat Rev Neurosci 12:509–523.

Jensen O, Mazaheri A (2010) Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front Hum Neurosci 4:1–8.

Jensen O, Bonnefond M, VanRullen R (2012) An oscillatory mechanism for prioritizing salient unattended stimuli. Trends Cogn Sci 16:200–206.

Klimesch W, Hanslmayr S, Sauseng P, Gruber WR, Doppelmayr M (2007a) P1 and traveling alpha waves: evidence for evoked oscillations. J Neurophysiol 97:1311–1318.

Klimesch W, Sauseng P, Hanslmayr S (2007b) EEG alpha oscillations: the inhibition – timing hypothesis. Brain Res Rev 53:63–88.

Lozano-Soldevilla D, VanRullen R (2019) The hidden spatial dimension of alpha: 10-Hz perceptual echoes propagate as periodic traveling waves in the human. Cell Rep 26:374–380.

Massimini M, Huber R, Ferrarelli F, Hill S, Tononi G (2004) The sleep slow oscillation as a traveling wave. J Neurosci 24:6862–6870.

Muller L, Reynaud A, Chavane F, Destexhe A (2014) The stimulus-evoked population response in visual cortex of awake monkey is a propagating wave. Nat Commun 5:3675.

Muller L, Piantoni G, Koller D, Cash SS, Halgren E, Sejnowski TJ (2016) Rotating waves during human sleep spindles organize

global patterns of activity that repeat precisely through the night. Elife 5:e17267.
Muller L, Chavane F, Reynolds J, Sejnowski TJ (2018) Cortical travelling waves: mechanisms and computational principles. Nat Rev Neurosci 19:255–268.
Nauhaus I, Busse L, Ringach DL, Carandini M (2012) Robustness of traveling waves in ongoing activity of visual cortex. J Neurosci 32:3088–3094.
Patten TM, Rennie CJ, Robinson PA, Gong P (2012) Human cortical traveling waves: dynamical properties and correlations with responses. PLoS One 7:e38392.
Petersen CCH, Hahn TTG, Mehta M, Grinvald A, Sakmann B (2003) Interaction of sensory responses with spontaneous depolarization in layer 2/3 barrel cortex. Proc Natl Acad Sci USA 100:13638–13643.
Pitts W, McCulloch WS (1947) How we know universals the perception of auditory and visual forms. Bull Math Biophys 9:127–147.
Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.
Sakata S, Harris KD (2009) Population activity in auditory cortex. Neuron 64:404–418.
Sato TK, Nauhaus I, Carandini M (2012) Traveling waves in visual cortex. Neuron 75:218–229.
Van Kerkoerle T, Self MW, Dagnino B, Gariel-Mathis MA, Poort J, Van Der Togt C, Roelfsema PR (2014) Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proc Natl Acad Sci USA 111:14332–14341.
VanRullen R (2016) Perceptual cycles. Trends Cogn Sci 20:723–735.
VanRullen R, Macdonald JSP (2012) Perceptual echoes at 10 Hz in the human brain. Curr Biol 22:995–999.
Varela FJ, Toro A, Roy John E, Schwartz EL (1981) Perceptual framing and cortical alpha rhythm. Neuropsychologia 19:675–686.
Watson AB, Pelli DG (1983) QUEST: a general multidimensional Bayesian adaptive psychometric method. Percept Psychophys 33:113–120.
Zhang H, Watrous AJ, Patel A, Jacobs J, Zhang H, Watrous AJ, Patel A, Jacobs J (2018) Theta and alpha oscillations are traveling waves in the human neocortex. Neuron 98:1269–1281.

2.3 Chapter Conclusion

Our results suggest that the ascending traveling waves only appear with the presence of bottom-up driven visual stimuli and disappear when visual inputs are absent; the descending waves, although they receive some modulation from external visual input, are less affected.

In accordance with the modeling study by Alamia and VanRullen (2019), oscillatory traveling waves might be a neural signature of predictive coding with forward waves carrying prediction errors and backward waves transmitting prediction signals. It bears mentioning that, in the empirical observation, predictions and error signals might mix together. Therefore, it might be inappropriate to directly link the observed biological signals with prediction or error signals suggested by predictive coding.

3 The biologically plausible neural network: deep predictive coding network

3.1 Chapter Introduction

In the second study, we evaluate the performance of predictive coding in deep neural networks. The idea is that if predictive coding can serve as the underlying processing principle in the brain, the model implementing predictive coding should behave similarly to humans to some extent. On the one hand, predictive coding has been modeled with the traditional method (see subsections 1.3.2.1 and 1.3.3.3). These models are concise and effective, but they can not handle large datasets and complex cognitive tasks like humans, which constrain the comparison between predictive coding models and human performance. Therefore it might be necessary to turn to deep neural networks powered by deep learning, as they are capable of simulating large numbers of neurons and parameters and thus capable of various complex tasks as reviewed in section 1.4.1.

On the other hand, deep neural networks face their own problems. For example, in computer vision tasks, different viewpoints of the same picture may result in completely different judgements by deep learning models. Could predictive coding help improve their performance? In recent years, the computational process of predictive coding has been implemented in deep neural networks (see section 1.4.2). These results show that predictive coding can bring better performance to deep neural networks. However, as one can see, those predictive coding models can vary in mathematical formulation, model construction, and training regimes. Which modeling method is the most likely to be adopted by the biological brain?

In the second study, we employed the model designed by Choksi et al. (2021). This model has been shown to have higher robustness towards noisy visual input. Does this model better reflect the way the brain operates? To answer this, one possible way is to test whether such a model could show human-like perception. Here we test the illusory perception of con-

tours. When humans are presented with illusory images, like a Kanisza square, in addition to the component inducers, an illusory square can also be perceived. Deep neural networks with high performance tend to report what they 'see', i.e., only the inducers. Here we test, whether the introduction of a predictive coding strategy into the deep neural networks could help gain the illusory perception.

3.2 Article 2

Predictive coding feedback results in perceived illusory contours in a recurrent neural network

Zhaoyang Pang, Callum Biggs O'May, Bhavin Choksi, and Rufin VanRullen. (2021) "Predictive coding feedback results in perceived illusory contours in a recurrent neural network." Neural Networks, 144. https://www.sciencedirect.com/science/article/abs/pii/S0893608021003373?dgcid=rss_sd_all

Neural Networks 144 (2021) 164–175

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Predictive coding feedback results in perceived illusory contours in a recurrent neural network

Zhaoyang Pang [a], Callum Biggs O'May [a], Bhavin Choksi [a], Rufin VanRullen [a,b,*]

[a] CerCO, CNRS UMR5549, Toulouse, France
[b] ANITI, Toulouse, France

ARTICLE INFO

Article history:
Available online 26 August 2021

Keywords:
Illusory contours
Predictive coding
Deep learning
Kanizsa squares
Feedback
Generative models

ABSTRACT

Modern feedforward convolutional neural networks (CNNs) can now solve some computer vision tasks at super-human levels. However, these networks only roughly mimic human visual perception. One difference from human vision is that they do not appear to perceive illusory contours (e.g. Kanizsa squares) in the same way humans do. Physiological evidence from visual cortex suggests that the perception of illusory contours could involve feedback connections. Would recurrent feedback neural networks perceive illusory contours like humans? In this work we equip a deep feedforward convolutional network with brain-inspired recurrent dynamics. The network was first pretrained with an unsupervised reconstruction objective on a natural image dataset, to expose it to natural object contour statistics. Then, a classification decision head was added and the model was finetuned on a form discrimination task: squares vs. randomly oriented inducer shapes (no illusory contour). Finally, the model was tested with the unfamiliar "illusory contour" configuration: inducer shapes oriented to form an illusory square. Compared with feedforward baselines, the iterative "predictive coding" feedback resulted in more illusory contours being classified as physical squares. The perception of the illusory contour was measurable in the luminance profile of the image reconstructions produced by the model, demonstrating that the model really "sees" the illusion. Ablation studies revealed that natural image pretraining and feedback error correction are both critical to the perception of the illusion. Finally we validated our conclusions in a deeper network (VGG): adding the same predictive coding feedback dynamics again leads to the perception of illusory contours.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

The human visual system is remarkably versatile. It is capable of accurately recognizing objects from different angles, over different distances, under distortion, and even under occlusion. Despite this, it can easily be fooled by simple visual illusions. Fig. 1 shows an example of Kanizsa illusory contour (or subjective contour) (Kanizsa, 1955, 1976). When observing such an image, most humans perceive the presence of a square, despite the only shapes present being the pacman-shaped inducers. In fact, in addition to perceiving edges of the illusory square, humans tend to perceive the interior of the induced shape as being brighter than the exterior, despite their being the same. These two phenomena illustrate two salient features of illusory figures — sharp illusory edges in regions of homogeneous luminance, and a brightness enhancement in the figure (Parks, 2001; Schumann, 1918; Spillmann & Dresp, 1995). There are many examples of illusions like this, demonstrating that the human visual system does not always accurately perceive stimuli. Such systematic misperceptions reflect underlying neural constraints and provide insight into the complex structure of visual cortex (Changizi et al., 2008). Visual illusions have thus been used as a probe for understanding visual processing (Eagleman, 2001; Gori et al., 2016). Alongside traditional neuroscientific approaches to studying visual perception, advances in machine learning have led to new avenues in understanding the mechanisms of visual processing. By developing artificial models of human vision we can investigate neuroscientific principles and reduce the need for experimentation. In particular, computational models allow for rapid, iterative experimentation and development. In fact, the origin of modern artificial neural networks (the backbone of much modern artificial intelligence or AI) is in computational models of neurons (McCulloch & Pitts, 1943). Throughout the history of AI development, researchers have taken inspiration from the brain. A clear example of this is the development of convolutional neural networks (CNNs), which were a critical step in the computer vision revolution (Fukushima & Miyake, 1982; LeCun et al., 1989). Inspired by the hierarchical structure of the brain, CNNs limit the spatial extent of the neuronal receptive

* Corresponding author at: CerCO, CNRS UMR5549, Toulouse, France.
E-mail address: rufin.vanrullen@cnrs.fr (R. VanRullen).

https://doi.org/10.1016/j.neunet.2021.08.024
0893-6080/© 2021 Elsevier Ltd. All rights reserved.

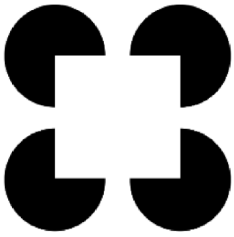


Fig. 1. Kanizsa square illusion.

fields, resulting in retinotopic feature maps much like the early visual system. In the last decade, the explosion of deep learning research has led to massive improvements in object recognition, even reaching super-human performance (He et al., 2015). However, the rapidly changing research landscape has produced many technical developments which are not always directly compatible with neuroscience. As a result, the relationship between human vision and computer vision has become less clear. A striking example of this is learning through error backpropagation, which uses a biologically implausible global error signal (Lillicrap et al., 2020). Recent attempts have been made to identify more biologically plausible learning rules (Ahmad et al., 2020; Lee et al., 2015; Millidge et al., 2020; Whittington & Bogacz, 2017). Many studies have also sought to investigate the similarities and differences between modern computer vision and human vision (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Wen et al., 2018b; Yamins et al., 2014). Such work highlights the dual benefits of computational neuroscience, which can both provide guidance for designing computer vision algorithms, and contribute to our understanding of brain functioning.

Despite all this, a recent report (Baker et al., 2018) shows that convolutional neural networks do not perceive illusory contours in the same way humans do. In their study, a feedforward CNN – AlexNet (Krizhevsky et al., 2012) – was trained to perform a thin/fat contour discrimination task on real and illusory contours. In both cases, the network could correctly classify the images, but the authors demonstrated that the representations of the illusory contours in the CNN do not resemble those of human observers. Specifically, they employed a so-called classification image technique (Gold et al., 2000) to give insight into which regions in a given image are important for the illusory formation. For human subjects, the region between inducers is critical to the perceptual decision (Gold et al., 2000). The feedforward neural network, on the other hand, failed to interpolate illusory contours between inducers — instead, it relied on the orientation of inducing elements to make its decisions (Baker et al., 2018). That is, instead of adopting a global processing strategy as humans did, the CNN mainly appeared to rely on local features. In a related study, Kim et al. (2021) directly examined the representations in intermediate layers of a feedforward neural network. The authors computed the cosine similarity between the network's representations of physical and illusory contours, and showed that they are more similar to each other than to control non-illusory shapes. However, this representation similarity measure is only an indirect indication of the perception of illusory contours, since it does not explain *which* features are similar between the illusory contours and the physical contours. It could be, for example, that the similarity is related to the presence of correctly-oriented corners, rather than to the contours between them. Thus, it may still be true that feedforward CNNs do not perceive illusory contours, but instead largely base their decisions on local features. Other evidence shows that feedforward CNNs can perform poorly on tasks which depend explicitly on global processing, like long-range spatial dependencies (Linsley et al., 2018) or object recognition under occlusion (Spoerer et al., 2017). In summary, it would seem that the visual perception of feedforward artificial neural networks is more dependent on local processing than human perception.

Perceptual discrepancies between artificial and biological networks highlight fundamental differences in their underlying structures as well as performed computations. Feedforward CNNs only roughly mimic the visual system, and particularly the feedforward pass through the visual pathway, while feedback connections, although abundant in the brain, are often ignored. Notably, physiological evidence shows that illusory contour perception may rely on feedback connections from V2 to V1 (Lee & Nguyen, 2001; Pak et al., 2020) or in higher-level regions (Cox et al., 2013; Pan et al., 2012). In this respect, the generation of illusory contours could be interpreted within the predictive coding framework (Notredame et al., 2014; Raman & Sarkar, 2016), a popular computational theory of feedback processing introduced in neuroscience by Rao and Ballard (1999). This theory posits that, in a hierarchical system, each layer tries to predict the activity of the layer below, and the prediction errors are used to update the activations. In fact, predictive coding could serve as a unifying computational principle for a variety of sensory systems including auditory (Kumar et al., 2011), olfactory (Zelano et al., 2011) and visual sensation (Mumford, 1992; Nour & Nour, 2015). In the field of machine learning, a number of recent works have also applied the predictive coding framework to modern deep learning networks. Boutin et al. (2021) demonstrated that a predictive coding network with a sparsity constraint learns similar receptive fields to neurons in V1 and V2, and provided evidence that the feedback helps the network perform contour integration. Lotter et al. (2017) showed that the predictive coding framework can be used to learn representations of pose and motion in video streams. Importantly, their model was only trained with unsupervised objectives, arguably closer to human learning than standard supervised learning methods. In addition to the biological plausibility of predictive coding networks, other works have demonstrated that these networks can show improved robustness to random or adversarial noise (Chalasani & Principe, 2013; Choksi et al., 2021; Huang et al., 2020). Wen et al. (2018a) studied whether predictive coding can result in improved classification of clean images, but since their network was not trained to minimize reconstruction error, it did not fulfil the predictive coding objective as described by Rao and Ballard (1999) (reduction of reconstruction errors over timesteps). A further discussion of the limits of the work of Wen et al. (2018a) can be found in Choksi et al. (2021).

Given both the significant role of feedback connections in biological illusory contour perception, and the potential for predictive coding as neuroscientific inspiration for feedback architectures in deep learning models, we hypothesized that a feedback neural network implementing predictive coding recurrent dynamics may perceive illusory contours in the same way humans do. The recent work of Lotter et al. (2018) strengthens this hypothesis. They investigated the responses of PredNet (Lotter et al., 2017), a predictive coding network, to illusory contours. They compared the response properties of model units to neuronal recordings in the primate visual cortex and showed comparable response dynamics in the presence of illusory contours (Lotter et al., 2018). Unlike the current work, which inspects the network's behavioural and perceptual responses, the Lotter et al. (2018) work approached the topic entirely at the level of neuronal activations. Although their research supports the idea of a similarity in contour representations between the PredNet artificial layers and the biological cortex, they do not read off network decisions, or inspect reconstructions of the network. Thus, they do not argue that the network is human-like on a behavioural

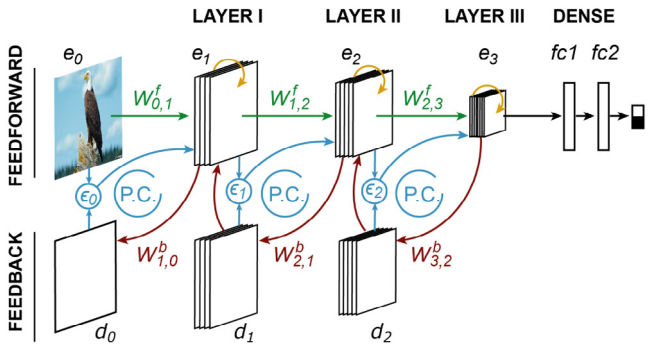


Fig. 2. Network architecture. The architecture consists of a main body and a classification head (or dense layers). For the main body, the predictive coding strategy is implemented in stacked autoencoders, with three feedforward encoding layers (e_n) and three generative feedback decoding layers (d_n). Reconstruction errors (ϵ_n) are computed and used for the proposed predictive coding updates which are denoted by "P.C." loops. Dense layers are added on top of the structure to implement a binary classification task.

Table 1
Table of parameters. Each encoding layer is a combination of a convolution layer and a ReLU nonlinearity with parameters Conv(channels, kernel size, stride). All convolutions have padding 4. Decoding layers consist of a deconvolutional layer with ReLU non-linearity, except for layer d_0, which uses a Sigmoid activation function, in order to compare with the input picture with pixel values ranging from 0 to 1. After flattening the output of the last convolutional layer, and going through a batch normalization function, two dense layers with a structure of weight (in features, out features) project to the binary decision layer.

Layers	Parameters
e_1	$\left[Conv(3,5,2)\right]_+$
e_2	$\left[Conv(128,5,2)\right]_+$
e_3	$\left[Conv(128,5,2)\right]_+$
d_0	$\left[Conv(3,5,2)\right]_{Sig.}$
d_1	$\left[Conv(128,5,2)\right]_+$
d_2	$\left[Conv(128,5,2)\right]_+$
$fc1$	$\left[W(2048,256)\right]_+$
$fc2$	$\left[W(256,128)\right]_+$

or perceptual level. As with the above-mentioned Kim et al. (2021) paper, it could be that the similarities in the representations are due to local low-level features (e.g. the inward-facing corners) rather than the illusory contours. Thus, it remains unclear whether their model truly "perceives" illusory shapes like humans do.

In the current study, we designed a deep predictive coding neural network according to the algorithm previously devised by our group (Choksi et al., 2021). We used a relatively small network, consisting of three stacked autoencoders, intended to roughly mimic the hierarchical structure of the early visual cortex in the primate brain, since neural correlates of illusory contours have been found to involve visual areas V1, V2 and V4 (Cox et al., 2013; Grosof et al., 1993; Von der Heydt et al., 1984; Pak et al., 2020; Pan et al., 2012). The network had both generative (image reconstruction) and discriminative (image classification) capabilities. Therefore, we could not only directly check whether the network indeed "sees" illusory contours, by examining its reconstruction images from generative feedback connections, but also measure the network's "behavioural" performance, i.e. the discriminative readout values indicating whether or not it detects an illusory shape. In addition, in order to equip our model with sufficient contour knowledge, we propose and verify that it is necessary to pretrain the neural network on natural images; this is in line with evidence that human contour integration and grouping is strongly tied to the statistics of the natural world (Geisler & Perry, 2009). Finally, we extend this approach to a deeper feedforward network (VGG) and demonstrate that this network sees the illusory contours too. In summary, we report that our predictive coding feedback networks tend to process illusory contours in a similar way to humans.

2. Materials and methods

2.1. Architecture

We construct a three-layer hierarchical stacked autoencoder with 3 feedforward encoding layers e_n ($n \in 1, 2, 3$) and 3 corresponding feedback decoding layers d_{n-1} (see Fig. 2 and Table 1). This is a conventional network architecture for learning a lower-dimensional latent representation of a high-dimensional input space (Kingma & Welling, 2014). When considering only the encoding layers, the network can be viewed as a standard feedforward convolutional neural network. To guide the implementation of the feedback connections, we follow the principles of "predictive coding" as introduced by Rao and Ballard (1999): in the hierarchical network, the higher layers try to predict the activity of the lower layers and the errors made in this prediction are then used to update their activity. For a given input image, we initiate the activations of all encoding layers with a feedforward pass. Then over successive recurrent iterations (referred to as timesteps t) we update the decoding and encoding layer representations using the following equations:

$$\begin{aligned} d_n(t) &= \left[W^b_{n+1,n} e_{n+1}(t)\right]_+ \\ e_n(t+1) &= \beta\left[W^f_{n-1,n} e_{n-1}(t+1)\right]_+ + \lambda d_n(t) \\ &\quad + (1-\beta-\lambda) e_n(t) - \alpha \nabla \epsilon_{n-1}(t) \end{aligned} \tag{1}$$

where $W^f_{n-1,n}$ denotes the feedforward weights connecting layer $n-1$ to layer n, and $W^b_{n+1,n}$ denotes the feedback weights from layer $n+1$ to n. $\epsilon_{n-1}(t)$ is the reconstruction error for layer $n-1$: the mean squared error between the representation e_{n-1} and the corresponding prediction d_{n-1}:

$$\epsilon_{n-1}(t) = \|e_{n-1}(t) - d_{n-1}(t)\|_2^2 \tag{2}$$

Then $\nabla\epsilon_{n-1}(t)$ denotes the gradient of the error at layer $n-1$ with respect to the activations in layer e_n. That is, $\nabla\epsilon_{n-1}(t)$ is the vector of partial derivatives of the error with respect to each element of e_n (and thus has the shape of e_n), so that the ith element of $\nabla\epsilon_{n-1}(t)$ will be $\frac{\partial \epsilon_{n-1}(t)}{\partial e_n(t)_i}$. We use the PyTorch automatic differentiation package to calculate this one-step gradient. The parameters β, λ and α act as balancing coefficients for the feedforward drive, feedback error correction, and feedforward error correction terms respectively, and they are treated as hyperparameters of the network (for the present experiments, except where otherwise noted, these hyperparameter values were fixed to $\beta = 0.2$, $\lambda = 0.1$ and $\alpha = 0.1$ as this was found to be sufficient for producing the illusion). We can also rewrite Eq. (1) by grouping the terms for each hyperparameter, to more clearly illustrate the connection with the Rao and Ballard formulation of predictive coding:

$$\begin{aligned} e_n(t+1) &= \beta\left[W^f_{n-1,n} e_{n-1}(t+1)\right]_+ + (1-\beta) e_n(t) \\ &\quad + \lambda(d_n(t) - e_n(t)) - \alpha \nabla \epsilon_{n-1}(t) \end{aligned} \tag{3}$$

This clarifies that the λ hyperparameter is the weight for the term $d_n(t) - e_n(t)$, which is exactly the gradient of $\epsilon_n(t)$ with

respect to $e_n(t)$. Thus we see that the λ and α terms are exactly the two error terms in Rao and Ballard's formulation. A fuller demonstration of the relationship between the two formulations (under the simplifying assumption of a linear activation function) can be found in the Supplementary Material.

Since the error ϵ_{n-1} is an average over the whole representation in the layer below (see Eq. (4)), as the number of units in the representation increases, the error term variance will tend to shrink to 0. Additionally, the retinotopic nature of the convolution operation (with connectivity restricted to a local neighbourhood) means that most connections between the layers are 0, and thus have 0 gradient. These combined effects result in small gradients relative to the layer's activations, so to counteract this we re-scale the gradient terms according to the layer and kernel sizes. Specifically, for each layer, we calculate for the layer below K = channels $\times$ width $\times$ height and C = channels $\times$ kernel size, and then multiplicatively scale the gradient term by a factor of $K/\sqrt{C}$. This directly balances out the two effects discussed. See the Supplementary Material for full details.

To reflect the systematic comparison between decoding and encoding layers, we set e_0 as our input images. Thus, our updating rule in Eq. (1) is only applied to $e_n (n \in 1, 2, 3)$ and e_0 will remain constant over timesteps. In addition, for the last layer e_3 in our model, there is no feedback, so we ignore the corresponding term in the update equation. All the weights W are fixed during the updates defined by Eq. (1). They are optimized over successive batches of natural images and across all timesteps (see Training procedure) to minimize the total reconstruction error L (Eq. (4))– an unsupervised objective in accordance with the principles of the predictive coding theory. In Eq. (4), N is the number of layers (here, $N = 3$). We note that Eq. (1) leads to updates which approximately reduce the loss in Eq. (4) over timesteps, as in both Rao and Ballard's formulation and Whittington and Bogacz (2017).

$$L = \sum_t \sum_{n=0}^{N-1} \epsilon_n(t) = \sum_t \sum_{n=0}^{N-1} \|e_n(t) - d_n(t)\|_2^2 \tag{4}$$

Intuitively, each of the four terms in Eq. (1) contributes different signals to a layer: (i) the feedforward term (controlled by parameter β) provides information about the (constant) input and changing representations in the lower layers, (ii) the feedback error correction term (parameter λ), hereafter referred to simply as "feedback", guides activations towards their representations from the higher levels, thereby reducing the reconstruction errors over time, (iii) the memory term helps to retain the current representation over successive timesteps, and (iv) the feedforward error correction term (controlled by parameter α) corrects representations in each layer such that their next prediction better matches the preceding layer, also contributing to the reduction in reconstruction errors. Together, the feedback and feedforward error correction terms fulfil the objective of predictive coding as laid out by Rao and Ballard (1999).

2.2. Training procedure

The network's training includes two stages: pretraining and finetuning (both using 10 timesteps for inference). The pretraining (over 150 epochs) was conducted in an unsupervised way with a reconstruction objective (see Eq. (4)), wherein both feedforward and feedback convolution weights were optimized over the CIFAR100 natural images dataset (Krizhevsky & Hinton, 2009). This was done to learn a hierarchy of relevant features to describe each natural image, as well as the corresponding generative pathway to reconstruct images from their features.

For the second stage, we added a 3-layer classification head to the network (consisting of three fully-connected layers), and finetuned all parameters of the network on the custom dataset presented below with a supervised binary cross-entropy loss. The weights of the whole network were finetuned for 25 epochs, after which it was tested with a new validation set. We performed three distinct pretrainings (each with different randomly initialized weights), and then each of these pretrainings was used to finetune 3 distinct networks. The reported test results are averaged over the resulting 9 networks.

During both pretraining and finetuning we use the backpropagation-through-time (BPTT) approach, which unrolls the network across all timesteps, and then backpropagates errors. We use the PyTorch automatic differentiation for efficiency and simplicity. We use the ADAM optimizer (Kingma & Ba, 2017), a stochastic gradient method with automatically-adapting learning rates. This algorithm automatically uses momentum (and we use the default PyTorch hyperparameters of 0.9, 0.999) to estimate the first and second order moments of gradients. We use an initial learning rate of $5e-5$.

When using BPTT, the computational requirements increase with the number of timesteps. Thus there is generally a balance required between the power of the model (where more timesteps allows for more complex computations) and feasibility. We chose 10 timesteps to produce meaningful dynamic trajectories while remaining within our computational constraints. We show in the Supplementary Material that we still see reasonable illusory perception results with other timestep values.

2.3. Stimuli

To test the perception of illusory contours we designed a custom dataset based on the Kanizsa illusion. We systematically generated stimuli consisting of an image with either a square, or four pacman-shaped inducers. The network was trained on a simple binary discrimination task: to classify each image as either a square, or pacman inducers. We generated four types of stimuli: (i) the real square; (ii) the illusory condition with all 4 inducers facing inwards (All-in); (iii) a control condition with all inducers facing outwards (All-out); and (iv) a random condition with random inducer orientation, but neither all in nor all out (Random). We finetuned the dataset for classification using only types (i) and (iv), and used (ii) and (iii) only at test time. In this way, the illusory shape is not part of the training set (out-of-sample), so the network cannot explicitly learn to categorise it either as a square or pacmen. The control condition was used to verify how the network reacts to stimuli which are out-of-sample but do not produce an illusion in humans.

Each image had a resolution of 32×32, consistent with the size of the CIFAR100 dataset used in the pretraining stage. The stimuli were designed to vary in luminance (for both the background and inducers, varying between 0 and 1), size, and position. This encourages the network to learn general rules and avoids overfitting through simple rote memorisation of the training set. For the same purpose, we also added Gaussian noise to each image (with variance randomly drawn from 7 pre-set levels varying from 0 to 0.3). Fig. 3 shows sample stimuli, with the different levels of noise. We generated a training set of 10,000 images, 5000 each for the Square and Random classes, and another set of 2500 images for validation. For testing, we generated 1200 images of each class for a total of 4800 images.

2.4. Feedforward baselines

The behaviour of the network at the first timestep (here $t = 1$, after feedforward initialization of activations throughout the network, but before error correction updates are applied according to Eq. (1)) can be understood as a pure feedforward network.

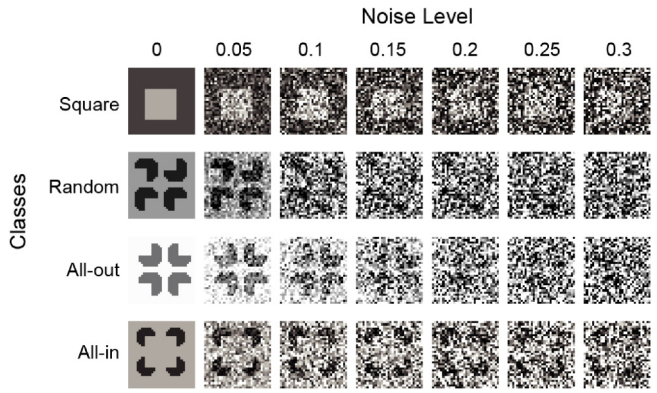


Fig. 3. Finetuning dataset. Sample training and testing images for 7 different levels of Gaussian noise: Square, Random, All-out, and All-in. All stimuli varied in luminance (both background luminance and inducer luminance), size (square sidelength or distance between inducers), and position. To enhance visualization, we displayed all seven levels of noises with each sample picture; while in practice, only one level of noise was randomly assigned to each stimulus configuration.

Table 2
Comparison of parameters for feedforward networks. PC is our predictive coding architecture. FF comprises only the feed-forward pass, but is trained with a supervised classification objective. FF-C is obtained by increasing the number of channels; FF-K by increasing the kernel size.

Models	Kernel size	No. channels	No. parameters
FF	5 × 5	(3,128,128)	1,403,620
FF-C	5 × 5	(3,172,172)	2,248,684
FF-K	7 × 7	(3,128,128)	2,199,268
PC	5 × 5	(3,128,128)	2,232,679

However, the network's features have been trained as part of the autoencoder blocks to optimize an unsupervised reconstruction objective — not (or not only) a supervised classification objective as in standard feedforward CNNs. We thus also chose to compare our model to a pure feedforward network (FF) trained only with a supervised classification loss. For a direct comparison, we also trained this network on CIFAR 100, before finetuning it on the shapes. This network has exactly the same architecture and number of parameters as the predictive coding network when considered only at $t = 1$, with the only difference being the learning objective.

However, this feedforward network has many fewer parameters overall than the full predictive coding network (around half). Thus, to directly compare to a feedforward network with roughly the same number of parameters, we also constructed two other feedforward networks: FF-C with more channels in each layer, and FF-K which uses a larger kernel size (Spoerer et al., 2017). Table 2 compares the number of parameters for each network.

2.5. Code availability

All the code, along with the pretrained models and stimulus datasets used in this work is available at: https://github.com/rufinv/Illusory-Contour-Predictive-Networks.

3. Results

3.1. Illusory contour perception

3.1.1. Shape classification

After pretraining on natural images and fine-tuning on simple shapes, the network could discriminate between physical squares and randomly oriented inducers in any configuration (except the two critical configurations, All-in and All-out, which had not been seen during training). During testing, all four classes were presented to the network, with All-out (control class) and All-in (illusory contours) as novel stimulus configurations formed by familiar inducers. Our hypothesis was that when presented with the All-in configuration, the network would assign a higher probability to the square class than when presented with the All-out configuration.

Table 3
Comparison to feedforward networks. The probability of square outcome for each class, for the feedforward baselines and the predictive coding network (including its initial timestep, which can also be viewed as a feedforward network, but with an additional unsupervised pretraining as part of a stacked autoencoder architecture). We see that the feedforward networks assign a square probability of almost 0 to all non-square classes, whereas the predictive coding network, by the last time timestep, has assigned a probability of nearly 0.5.

Models	Square	No. random	All-out	All-in
FF	0.999	0.000	0.002	0.002
FF-C	0.999	0.000	0.002	0.003
FF-K	0.999	0.000	0.002	0.003
PC(t=1)	0.981	0.000	0.001	0.000
PC(t=100)	1.000	0.000	0.201	**0.478**

For each class, we thus inspected the probability assigned by the network to the square category for each image (Fig. 4, results averaged over 9 networks). At timestep 1 (with only feedforward processing having taken place), the network appears to classify images based on low-level, local information, as all the pacmen-made patterns (Random, All-out, All-in) are recognized as non-square shapes. However, over timesteps, the network begins to recognize the All-in condition (the illusory contour) as a square, at a much higher rate than the All-out and Random conditions. After 50 timesteps, the average probability assigned by the network to the square class (an "illusory contour perception") increases by more than 40% for the All-in condition, compared to around 20% (depending on noise level) for the other conditions. Although this measure does not go above 50%, we suggest that even humans would not actually categorize inducers as squares—despite "seeing" the illusion, we easily recognize that there is no actual square in the image. As the noise level increases (Fig. 4, right), the likelihood of reporting a square for the All-in condition stays roughly constant, whereas the likelihood for the All-out and Random conditions increases. This can be interpreted as the network becoming more 'confused', as all probabilities are drifting together. The *difference* between the illusory condition and the control conditions gets smaller as the noise level increases, but remains visible for high noise. This is consistent with previous work which demonstrated that humans perceive illusory contours even under noise (Gold et al., 2000).

We also compared our predictive coding network to the various feedforward baselines. Table 3 shows the results. We clearly see that all feedforward networks assign a near-zero average square probability to both the All-in and All-out conditions, showing that they do not perceive the illusion at all. This further confirms the hypothesis that feedback connections are critical for the perception of illusory contours.

3.1.2. Image reconstructions

Although the network assigned higher probability of square to the illusory contours than to the control condition, it remains hard to draw the conclusion that the network could really "see" illusory contours. It could still be the case that the network is basing its classification decision on other features (e.g. the correctly-oriented corners). A major advantage of the predictive coding model is that we can use its generative feedback pathway to inspect the image reconstructions produced by the model. We

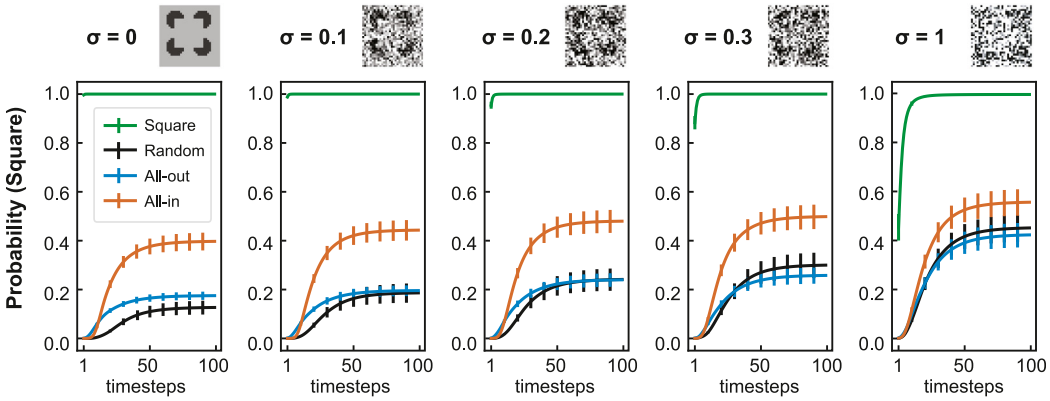


Fig. 4. Classification results. The probability of "square" report over recurrent predictive coding iterations (timesteps). Each panel shows a different noise level — for visibility, only three out of six levels of noise are shown here, in addition to the clean images ($\sigma = 0$). Feedback iterations increase the likelihood of "square" report, especially for the All-in (illusory contour) condition. Results are averaged over 9 networks and error bars represent standard error of the mean (SEM) across the networks.

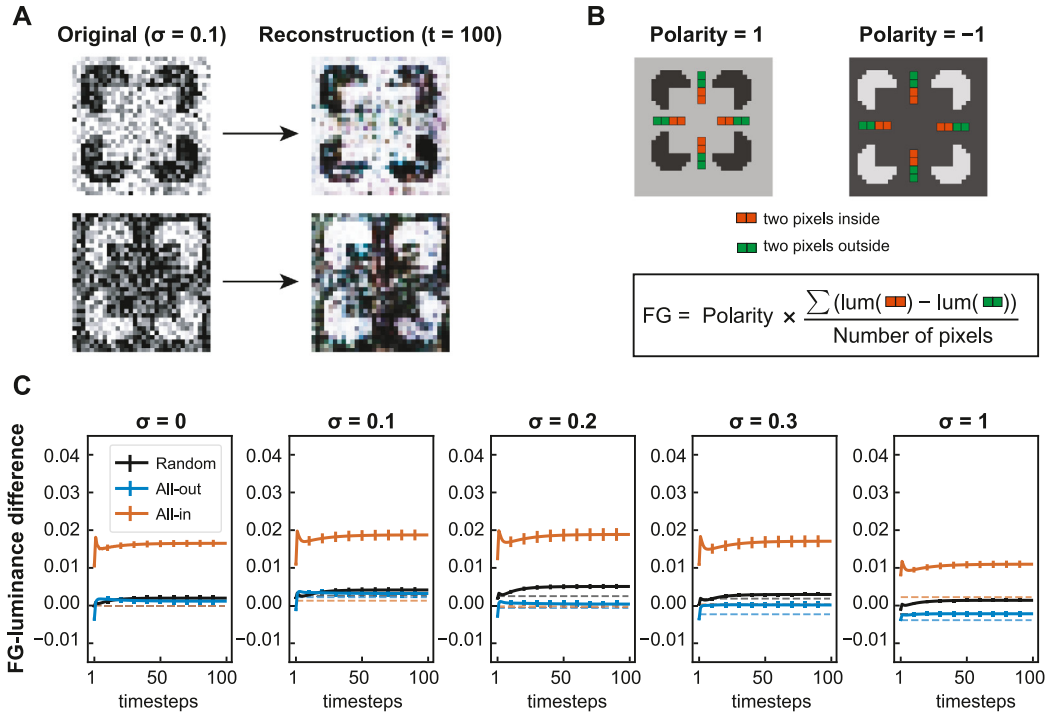


Fig. 5. Quantification of illusory contour perception in the neural network. A. Two examples of illusory contour (All-in) stimuli, and their corresponding reconstructions from the network at timestep 100. B. Computation of the "FG" value, measuring the figure-ground luminance difference. C. FG-luminance difference for classes Random, All-out, and All-in over 100 timesteps (zero is absence of illusion, larger values mean more illusion is perceived). Dashed lines denote the FG values computed from original input pictures of the corresponding classes. With clean images, we expect FG on input pictures to be equal to zero as shown in the first plot, while FG values can be bigger or smaller than zero with noisy input images. Results are averaged over 9 networks (distinct random initializations, pretraining and fine-tuning) and error bars represent SEM across networks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

inspect the reconstruction of the bottom layer of the network (that is, $d_0(t) = W^b_{1,0} e_1(t)$). Fig. 5A displays two examples of reconstructed images at timestep 100. Compared to the original images (on the left) whose noise standard deviation is 0.1, the reconstructed images are denoised by the network, and the illusory contour shapes appear clearer. (Note that this is an emergent property of predictive coding, and not a built-in property of our training scheme: the network's body was not trained with a denoising objective, but only to reconstruct clean natural images). To quantify the illusion "perceived" by the network, we examined the luminance profile of the reconstructions: for each image we computed a "Figure-Ground luminance difference" (FG), as illustrated in Fig. 5B. This statistic measures whether the network perceives an illusory brightness enhancement like humans. Given the expected position of the illusory contour (the position that the square would have occupied if it had been real instead of illusory), along each of the four cardinal axes we took two pixels inside (red in Fig. 5B) and two pixels outside the square (green in Fig. 5B). We computed the average difference between the pixel luminance values inside and outside. A polarity factor (-1 or 1) multiplied this measure, to take into account the different configurations: dark inducers (polarity = 1) are expected to produce lighter illusory shapes, light inducers (polarity = -1) to produce darker ones. The constructed FG measure is zero in the original All-in images (since it is measured in the background between the inducers), but should be positive in the image reconstructions

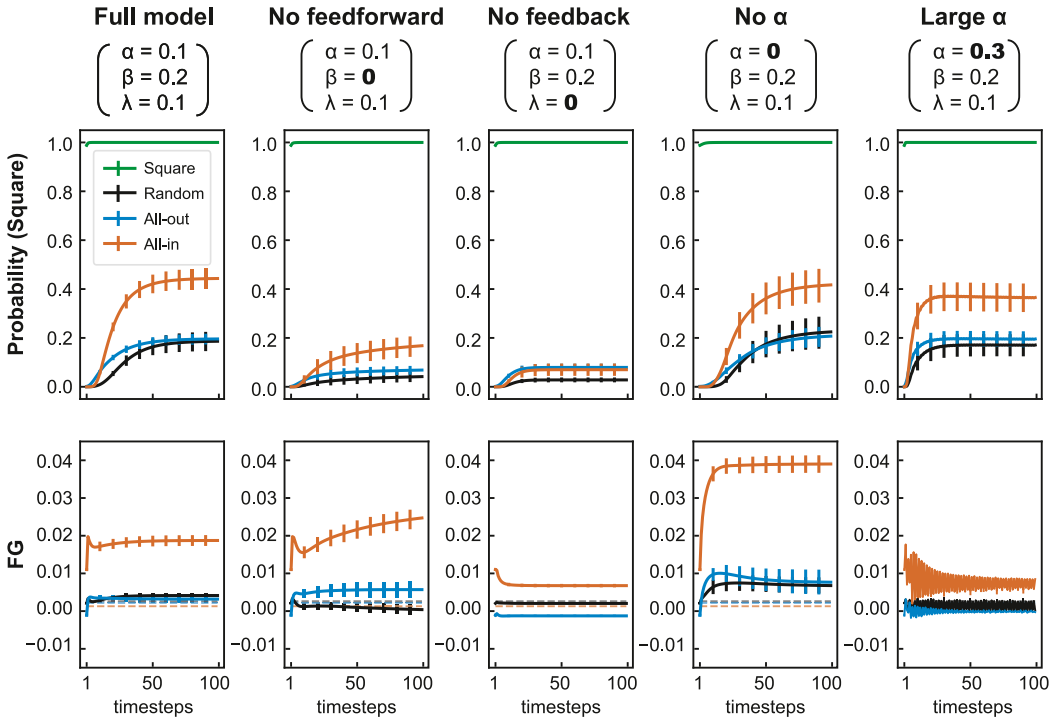


Fig. 6. Ablation during testing. The full model results (first column) correspond to the same data already reported in Figs. 4 and 5 for $\sigma = 0.1$. Probability of square classification is reported in the top row, FG value in the bottom row. Each component of the update equations was set to 0 at test time (with the other parameters at their default value): the feedforward drive term β (2nd column), the feedback term λ (3rd column), the feedforward error correction term α (4th column). In addition, we tested a larger value of α (last column).

whenever an illusory contour is perceived. Fig. 5C shows how this FG value changes over predictive coding iterations, for illusory contours (All-in class) and the other two non-illusory shapes (All-out and Random classes). For the All-in class, after the initial time step the value is consistently higher than zero, and 5 to 10 times higher than the FG value measured for the control (All-out) or the random inducer classes. The corresponding luminance difference is small but reliable (on the order of 0.02, with 1 denoting the full luminance range). For comparison, the same FG value for the physical square would range from 0.3 to 0.6 (not shown here). Interestingly, as seen above, this perception of illusory contours was still present but somewhat reduced when images were corrupted with extremely high levels of noise ($\sigma = 1.0$), a behaviour that would be expected in humans. For all other noise-levels that the network was trained to reconstruct, the results are relatively similar; thus, from here onward we only plot results for a single level of noise (0.1), even though all our conclusions equally apply to all levels.

3.2. Ablation studies

In order to determine the contribution of the various components of the network to illusory contour perception, we performed systematic ablation studies. For each of the hyperparameters α, β and λ in the update Eq. (1), we set the parameter to 0 at test time and observed how this affects the network behaviour. Additionally, we also tested a network with a larger value of α, keeping all other hyperparameters at their default value. Fig. 6 shows the results of these ablation experiments.

As expected, removing feedback (λ) leads to the complete disappearance of the illusion. This confirms that the generative feedback plays a critical role in producing the illusion. Interestingly, we see that removing the constant feedforward drive β (after the initial feedforward activation pass) also seems to diminish the illusory effect, although not completely. It thus appears that both feedforward and feedback contributions are important to see the illusion. Finally, when removing the feedforward error correction term α, the square classification probability does not seem to change much; however, the FG values strongly increase, specifically for the All-in condition. This suggests that the network may be seeing the illusion even more strongly. This can be explained by the fact that the role of the feedforward error term α is to update the activations in each layer in order to improve their reconstruction of the layer below. Thus, large FG values (in fact reflecting an imperfect reconstruction of the input image caused by the illusion) are "corrected" slightly by the feedforward error correction. When α is set to 0, the network does not correct these large FG values anymore, and is free to "perceive" a stronger illusion. To verify that feedforward error correction does indeed suppress the illusory contour perception, we also tested a model with stronger α. In this case, both the square classification probability and FG value are visibly decreased for the All-in condition, confirming our hypothesis.

In Rao and Ballard's original formulation (Rao & Ballard, 1999), they treated the two error terms (feedforward and feedback error corrections) together, and did not explore their relative effects. Here we show that only one of these (the feedback error correction) is critical to the perception of illusory contours, but that overall this perception is consistent with the predictive coding framework. However, the network with $\alpha = 0$ is essentially equivalent to a regular feedback network, so this suggests that other feedback architectures may also lead to the perception of illusory contours.

3.3. Training restricted networks

To further explore the effects of the various network components, we also trained networks without each component. The results are shown in Fig. 7. The difference from the previous ablation studies (Fig. 6) is that here, rather than being

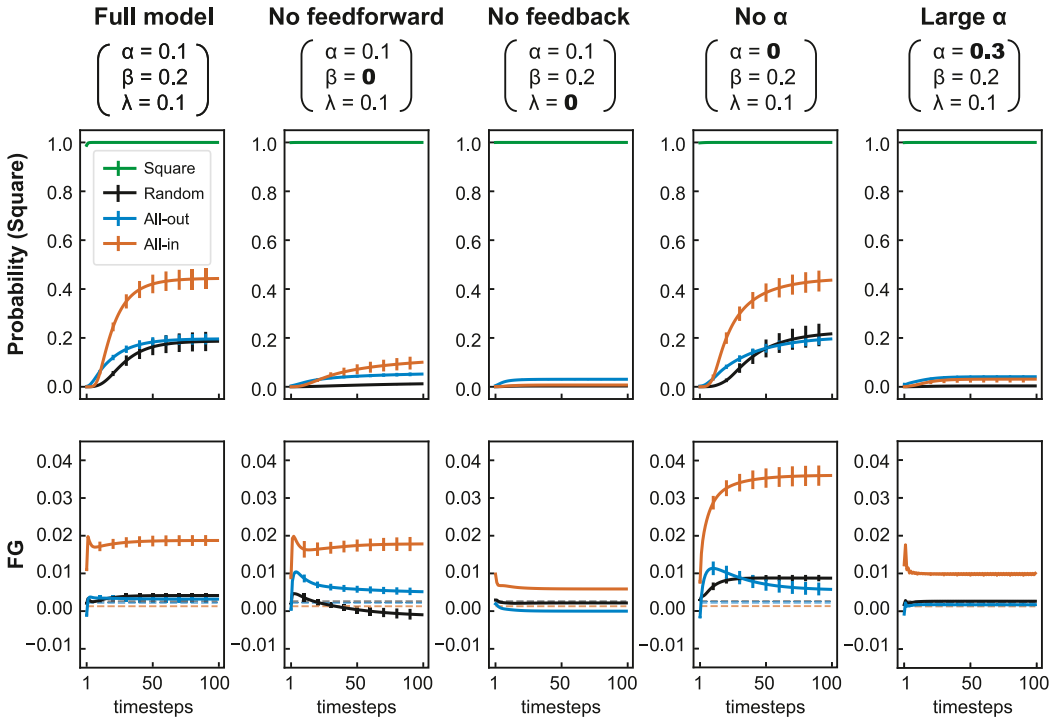


Fig. 7. Training restricted networks. These graphs show the behaviour of networks trained from scratch without the various components of the full model. The full model results (first column) correspond to the same data already reported in Figs. 4 and 5 for $\sigma = 0.1$. Probability of square classification is reported in the top row, FG value in the bottom row. Generally, this produces similar results to just switching off the components at test time. Again, feedback is essential to the perception of the illusion, and feedforward encourages it. On the other hand, α being larger tends to diminish the perception of the illusion.

trained with all the components and then having them artificially switched off at test time, here the network "knows" during training which components are unavailable, and can potentially learn to compensate for their absence. Thus, where the previous section answered the question "How does our full trained network behave without each component", this approach answers the question "What is the behaviour of the best possible model without each component". That being said, overall this experiment largely confirms the results of the above ablation tests: that the feedback is essential for the perception of the illusion; that feedforward significantly contributes to this perception; and that larger α (feedforward error correction) diminishes this illusion. One significant difference between Figs. 6 and 7 is that training with a large α seems to have a stronger deleterious effect than just increasing it at test time. We speculate that this is due to the fact that when the network is trained with large α, it implicitly learns to "trust" the reconstructions from higher levels more (since they will be corrected faster), and thus learns larger feedback weights.

3.4. Pretraining and finetuning datasets

We also investigated to what extent the specific datasets used for pretraining and/or fine-tuning the network affect its behaviour. Fig. 8 shows the results of three comparative tests. In the first test, we took our three pretrained networks (without finetuning on the custom shapes dataset) and tested whether they can "perceive" illusory contours solely based on the learned statistics of natural images (therefore, this test relies only on the FG figure-ground luminance calculation). In the second test, to further confirm the critical role of *natural images* (instead of other much simpler but task-related images) in illusion perception, we simply removed the pretraining procedure on the CIFAR100 natural image dataset. That is, we trained our network directly on the shape dataset (for both unsupervised reconstruction pretraining, and supervised classification finetuning). In the third test, we trained without any noise during the finetuning (supervised classification) stage.

Figs. 8B and C confirm our hypothesis that pretraining on natural images is necessary and sufficient for illusion perception. For Fig. 8B, we can only measure the FG values from image reconstructions, but not the classification probabilities, since the pretrained networks were not equipped with a decision head at this stage. Figure S1-A in the Supplementary Material shows that with the default set of parameters, the pretrained networks can already perceive a certain amount of illusion. Here, to maximize the networks' ability to perceive illusory contours, we applied a different set of parameters with $\beta = 0.1$, $\lambda = 0.2$ and $\alpha = 0$ (since we have observed in our ablation tests above that increasing λ and decreasing α could favour the illusion; default parameters are still used elsewhere unless stated otherwise). We also test these networks on clean shapes, because they were only trained on the clean CIFAR100 natural images. From Fig. 8B, it is clear that pretraining on natural images can provide sufficient knowledge about image contour statistics to induce the perceptual illusion. Although the magnitude of the effect is somewhat dependent on the choice of parameters, as detailed above and in our ablation studies (Figs. 6 and 7), in a separate simulation (Figure S1-B) we show that illusory contours can even arise in a parameter-free alternative training regime (similar to the one described below in 3.5). In this case, feed-forward and feedback connections are trained to optimize reconstruction after a single timestep—so the parameters α, β and λ play no role in the outcome. Yet the networks could still perceive illusory contours (Figure S1-B). Fig. 8C shows that when the network convolutions are only trained to reconstruct shapes (squares, pacman inducers) instead of natural images, the model processes the All-in and All-out test images similarly. Both results provide evidence that the perception of illusory contours is an emergent

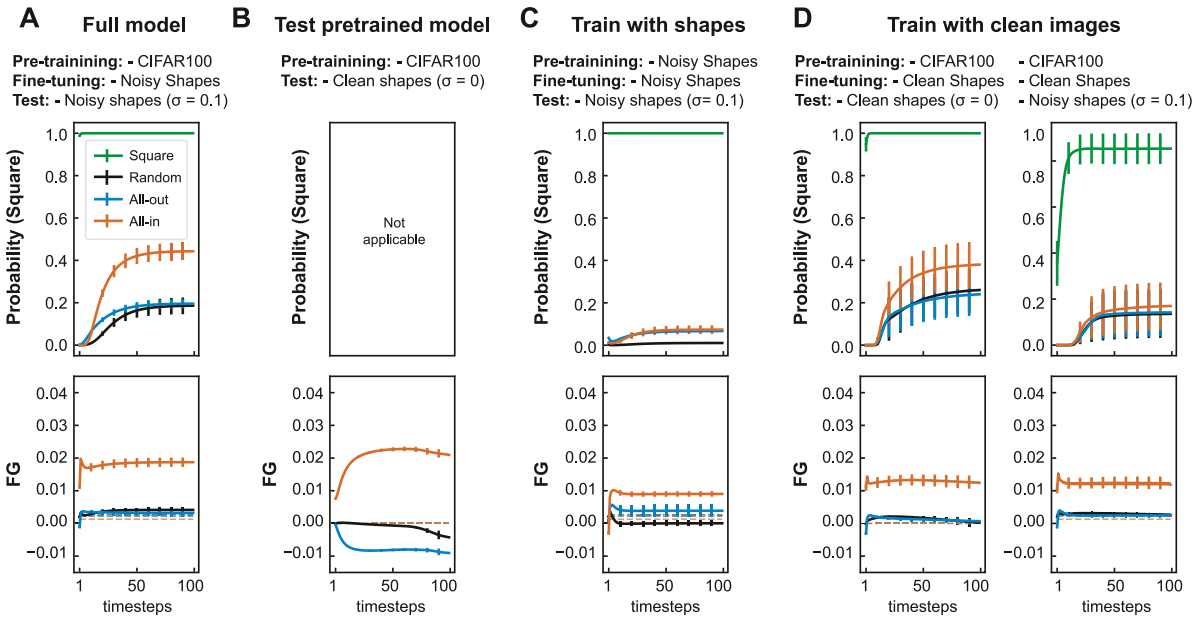


Fig. 8. Testing the influence of pretraining and finetuning datasets. A. The full model is the same data already reported in Figs. 4, 5, 6 and 7, to facilitate comparisons. B. Testing whether networks that have only been pretrained on natural images (no fine-tuning) could perceive illusory contours (this test relies on FG values, as classification probability is not computable for these networks) C. Comparing pretraining on CIFAR100 (the Full model) to training directly on the shapes dataset. D. Comparing finetuning on a noisy shape dataset (the Full model) to a clean alternative, i.e. trained and finetuned without noise. The testing is performed without noise (left) or with $\sigma = 0.1$ noise (right).

property from the visual system's adaptation to the statistics of the natural world. The third test investigates the effect of noise (and therefore, uncertainty) during shape classification training. Fig. 8D demonstrates that training without any noise results in a weaker illusory effect for both the square class probability and the FG value. That is, some form of uncertainty or variety in the learned shapes dataset appears helpful; otherwise the network may find a way to perfectly encode or memorise the stimuli, i.e. "overfit" the training set.

3.5. Illusory contour perception in a modern deep network equipped with predictive coding dynamics

Finally, we asked whether our approach – equipping a feedforward network with recurrent predictive coding dynamics – could be applied to other networks to produce illusory contour perception. Here we used VGG, a much deeper 11-layered network (Simonyan & Zisserman, 2015), which more closely resembles modern state-of-the-art computer vision networks. Previous work from our group demonstrated that the VGG architecture can be equipped with predictive coding dynamics (Choksi et al., 2021), and we follow the same approach here.

The basic VGG architecture is a simple feedforward CNN with 8 convolutional layers with 3×3 kernels, max pooling, and 3 fully-connected layers for the head. Although now a few years out of date, it was state-of-the-art in 2014. Compared with the original architecture that was designed for larger images from the Imagenet dataset (Simonyan & Zisserman, 2015), the main difference for our network trained on smaller natural scenes (32×32 pixels) is that the first pooling layer was removed, and the final classification output layer was changed to have 100 instead of 1000 classes. We use a feedforward backbone network where the weights were optimized for natural image classification on the CIFAR100 dataset. The trained model is then augmented with predictive loops consisting of transpose convolution layers, as in the 3-layer model. However for this larger model, the feedback predictions span two feedforward convolutional layers, rather than just one. Thus the inputs to the feedback convolutions are the 2nd, 4th, 6th and 8th convolutional layers (respectively aiming to reconstruct the image input, 2nd, 4th and 6th layers). The weights of these feedback convolutions are optimized for one-step reconstruction over CIFAR100 with a simple mean square error loss (while the feedforward weights remain fixed). Finally the same update equations (Eqs. (1) and (4)) are used as with the 3-layer model. This recurrent model is hereafter referred to as PVGG (for "predictive-VGG"). Notably, the feedback connections in the PVGG model were trained with a different method (already alluded to above), optimizing image reconstruction on the CIFAR100 dataset after a single timestep. Therefore, the parameters α, β and λ play no role in the training. Then, during testing, we selected a set of parameters according to our observations with the smaller model. As we found that the feedforward error correction term α tends to suppress the illusion and the feedback λ tends to increase it (Figs. 6 and 7), we here set $\alpha = 0$, $\beta = 0.1$, $\lambda = 0.2$.

Since we do not train the model at all on the shapes dataset, we cannot examine the network's classification decisions (since in order to do so, at the very least the classification head would need to be re-trained with the appropriate shape classes). Instead we directly inspected the image reconstructions of the network and calculated the FG values, as illustrated in Fig. 5. Fig. 9 shows the results: it is clear that over timesteps, once again the average FG value for the illusory inducers rapidly becomes much larger than either the random or control conditions. As in previous experiments, we could verify that feedback error correction is critical for this illusory perception: no illusion occurred when we ablated the feedback at test time, i.e. $\lambda = 0$ (Fig. 9, right).

Although this network also appears to perceive the illusion, the long-run dynamics of the network are quite different from the smaller model. While in the smaller model, even after 100 timesteps the illusion is still perceived, in the PVGG model it gradually disappears after some timesteps. In a way, presenting a static stimulus for so many timesteps is somewhat unrealistic — for humans, the input stimulus is constantly changing, either because of environmental fluctuations, or because of saccadic eye movements. Indeed, previous work showed that when illusory

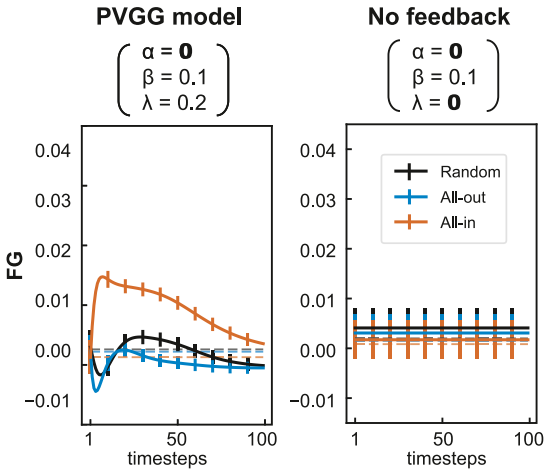


Fig. 9. FG values in the PVGG network. Left: The average FG value for the inducer condition is significantly larger than for the two control conditions, demonstrating that the network perceives the illusion. Right: Removing feedback (which here leads to constant activations, since there is no feedforward error correction either) clearly destroys the illusion.

contours are perceived by humans in the periphery, they disappear after a few seconds (Ramachandran et al., 1994). However, the difference between PVGG and the smaller model is interesting, and could be an avenue for further research. A number of key differences between the two models could be at play. Most importantly, the weights of PVGG are not learned "over timesteps" - the feedforward weights are trained just for feedforward classification, and the feedback weights are trained just for a single step reconstruction. Thus, the network never updates its weights after predictive coding updates. On the one hand, this makes it even more remarkable that the updates lead to the perception of the illusion, and is strong evidence that the framework is compatible with illusory contour perception. However, it could also be the reason why the long-term dynamics are somewhat unstable (or non-convergent), unlike in the smaller model. The other significant differences are that PVGG is a much deeper network, and was trained only on natural images, never on the shapes dataset. As such it has much more complex, abstract representations, and it is possible that these representations are not well-calibrated to simple shapes or contours which would not appear in CIFAR images. As a result the network might struggle to maintain stable representations of the illusory contours, and might instead "hallucinate" (or "fill in") what it imagines is missing texture or complexity.

The above strategy of avoiding re-training or fine-tuning the network has advantages from both an AI and a neuroscience perspective. First, not only does it save the time and power required for training, but it also allows us to experiment with any pretrained feed-forward model — including models which we would not be able to train ourselves due to computational restrictions. Second, from a neuroscience perspective, the fact that the perception of the illusion can occur in a network which has never seen the shapes dataset before (nor, presumably, any square or pacman shape) demonstrates even more clearly that it results from a combination of feedback connectivity, predictive coding dynamics, and exposure to natural scenes.

4. Discussion

The purpose of this study was to test whether a feedback neural network with brain-inspired recurrent dynamics would perceive illusory contours (Kanizsa squares) in a similar manner to humans. Augmenting a feedforward CNN with predictive coding recurrent dynamics allowed us to (i) analyse explicit classification decisions (square vs. inducers) and, unlike other related work (Baker et al., 2018; Kim et al., 2021; Lotter et al., 2018), (ii) visualize reconstructed inputs from the model's viewpoint. As reported in a preliminary version of this study recently published in a conference workshop (Pang et al., 2020), we found that, compared to a feedforward baseline, the recurrent dynamics led the network to perceive more illusory contours. Notably, by inspecting the network's reconstructions, we were able to directly visualize the network's internal representation of the stimulus, which provides a much clearer measure of "illusory perception" than previous works. We found evidence of modulations of the perceived luminance profiles in the expected direction for illusory shapes, suggesting that the network is truly "perceiving" the contours. We extended this analysis and performed systematic ablation studies, both at test time and at training time, and found that the feedback error correction term is essential to the perception of the illusion, while the feedforward error correction term tends to decrease it. Similarly, exploring the datasets used for pretraining and fine-tuning the network revealed that prior exposure to the statistics of natural scenes is a crucial element of illusory contour perception. Finally, we also implemented the predictive coding dynamics in a standard VGG model, and found that the modified PVGG model also exhibited the perception of illusory contours. This suggests that illusory contour perception arises from predictive coding feedback dynamics, independently of the scale of the model. In summary, we provide clear evidence that brain-inspired recurrent dynamics can lead networks to perceive illusory contours like humans.

Although there are intrinsic differences between the human visual system and artificial neural networks (e.g., the global error signals required for learning via back-propagation (Lillicrap et al., 2020)), we argue that the current findings highlight three key similarities with biological vision. First, both systems may engage similar global processing of illusory contours. In the visual system, Pan et al. (2012) reported that illusory contours activate equivalent representations in V4 compared to real contours, whereas V1 and V2 differently encode their respective local features. That is, in addition to local processing in early visual regions, there exists a global processing mode whereby illusory inducers form integral contour representations. In a similar way, when presented with illusory contours the current network assigned much higher probability of the "square" class than for either Random or All-out control images, though they shared the same local features as the All-in illusory contour images. This indicates that the network also possesses a capability of global processing. Moreover, this global processing primarily results from the feedback connections, since none of the tested feedforward networks could perceive illusory contours (Table 3). Second, the "behavioural" performance (i.e. decision probability) of the network is also consistent with physiological research on illusory contours. Lee and Nguyen (2001) compared EEG activity for illusory contours and other patterns, and found that the activity for illusory contours is significantly higher than control random stimuli, but still lower than real contours. In the current study, the "Square" class probabilities assigned by the network after the Softmax layer indicate a similar pattern (see Fig. 4). Lastly, at the "perceptual" level, we directly checked the internal representation at the first layer of the network (through its generative "image reconstruction" pathway). The FG metric suggested that the network perceives a brighter (or darker) illusory shape, consistently with the "illusory brightness" reported when humans perceive illusory contours (Parks, 2001; Schumann, 1918; Spillmann & Dresp, 1995).

Having designed a biologically inspired architecture, we performed ablation studies to examine how the behaviour of the network depends on its various components. Most importantly, we found that feedback error correction was critical to the network perceiving the illusory contours, which is in agreement with previous results which demonstrated that feedforward networks do not appear to perceive illusory contours (Baker et al., 2018). We also found that removing the feedforward error correction term (one half of the predictive coding update proposed by Rao and Ballard (1999)) seems to enhance the perception of the illusion. This is easily explained as this term serves to correct the representations to minimise reconstruction error of the incoming layer, so the erroneous contour and luminance difference may be "corrected out" when this term is present. These ablation experiments allow us to highlight how the two error correction terms, introduced concurrently in Rao and Ballard (1999), play two distinct roles: the feedforward error correction term encourages the network to accurately represent the stimuli, anchoring it in the ground truth; whereas the feedback error correction tries to explain the input in terms of the network's implicit priors, and in this case causes the network to "hallucinate" based on its higher-level representations (for example, closed contours or square shapes). Finally, by scaling the predictive coding updating dynamics to VGG11 and finding the same human-like illusory contours (see Fig. 9), we provided evidence that these feedback dynamics may induce illusory perception across a large range of deep convolutional network architectures.

In Section 3.4, we investigated what effect pretraining on natural images has on the network. Previous work suggested that CNNs trained on natural scenes are better models of brain activity (Maheswaranathan et al., 2018). Indeed, we saw that training directly on the shapes was not sufficient for the network to assign higher probability of the square class to illusory contours, nor to produce the positive FG values which are the hallmarks of illusory contour perception. Training on natural images encourages the network to learn efficient representations of natural stimuli — the constraints that arise from exposure to natural scene statistics in the brain have been proposed as a cause of many visual illusions (Eagleman, 2001; Gori et al., 2016). In other words, pretraining the network on natural images may help the network to learn more human-like representations. We also investigated the effect of training with images corrupted with Gaussian noise versus without. Such data augmentation typically helps in improving generalisation as it forces the network to learn meaningful features from the data instead of rote-memorising individual training examples (Akbiyik, 2020; Bishop, 1995). We found that networks trained using noisy images perceive stronger and more consistent (as reflected by smaller standard error values) illusory contours.

In summary, by leveraging insights from neuroscience, we designed an original brain-inspired deep learning architecture and thus add to a growing body of literature exploring the cross-pollination of neuroscience and AI. On the one hand, the current study demonstrates that we can effectively use neuroscience principles to design artificial computer vision models, and probe them using classical stimuli and illusions from neuroscience and cognitive science. On the other hand, we also illustrate how modern deep learning techniques can be used as powerful tools in examining our theories of brain function (Cichy et al., 2016; Kriegeskorte, 2015; Marblestone et al., 2016; VanRullen, 2017). By building and testing a brain-inspired model, the current study highlights the essential roles of feedback connections (Lee & Nguyen, 2001; Pak et al., 2020), predictive coding computation (Notredame et al., 2014; Nour & Nour, 2015; Raman & Sarkar, 2016; Shipp, 2016), and prior experience of natural environments (Eagleman, 2001) for the perception of visual illusions. Future work could use the same kind of predictive coding model to test other aspects of the predictive coding theory in neuroscience, such as its tendency to produce oscillatory dynamics (Alamia & VanRullen, 2019), or other phenomena observed in human vision, such as ambiguous stimuli and multi-stable perception, or the Gestalt rules of perceptual organization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Milad Mozafari for providing the implementation for PVGG. This work was funded by an ANITI (Artificial and Natural Intelligence Toulouse Institute), France Research Chair to RV (ANR grant ANR-19-PI3A-0004), as well as ANR, France grants AI-REPS (ANR-18-CE37-0007-01) and OSCI-DEEP, France (ANR-19-NEUC-0004). ZP is supported by China Scholarship Council (201806620059).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2021.08.024.

References

Ahmad, N., van Gerven, M. A. J., & Ambrogioni, L. (2020). GAIT-prop: A biologically plausible learning rule derived from backpropagation of error. arXiv:2006.06438.

Akbiyik, M. E. (2020). Data augmentation in training CNNs: Injecting noise to images. URL: https://openreview.net/forum?id=SkeKtyHYPS.

Alamia, A., & VanRullen, R. (2019). Alpha oscillations and traveling waves: Signatures of predictive coding? *PLOS Biology*, *17*(10), 1–26. http://dx.doi.org/10.1371/journal.pbio.3000487.

Baker, N., Erlikhman, G., Kellman, P. J., & Lu, H. (2018). Deep convolutional networks do not perceive illusory contours. In *Proceedings of the 40th annual conference of the cognitive science society*. Madison, WI: Cognitive Science Society.

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, *7*(1), 108–116.

Boutin, V., Franciosini, A., Chavane, F., Ruffier, F., & Perrinet, L. (2021). Sparse deep predictive coding captures contour integration capabilities of the early visual system. *PLoS Computational Biology*, *17*(1), Article e1008629.

Chalasani, R., & Principe, J. C. (2013). Deep predictive coding networks. arXiv:1301.3541.

Changizi, M. A., Hsieh, A., Nijhawan, R., Kanai, R., & Shimojo, S. (2008). Perceiving the present and a systematization of illusions. *Cognitive Science*, *32*(3), 459–503.

Choksi, B., Mozafari, M., O'May, C. B., Ador, B., Alamia, A., & VanRullen, R. (2021). Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. arXiv:2106.02749.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Cox, M. A., Schmid, M. C., Peters, A. J., Saunders, R. C., Leopold, D. A., & Maier, A. (2013). Receptive field focus of visual area V4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences*, *110*(42), 17095–17100.

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, *2*(12), 920–926.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets* (pp. 267–285). Springer.

Geisler, W. S., & Perry, J. S. (2009). Contour statistics in natural images: Grouping across occlusions. *Visual Neuroscience*, *26*(1), 109–121. http://dx.doi.org/10.1017/S0952523808080875.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, *10*(11), 663–666.

Gori, S., Molteni, M., & Facoetti, A. (2016). Visual illusions: An interesting tool to investigate developmental dyslexia and autism spectrum disorder. *Frontiers in Human Neuroscience*, *10*, 175.

Grosof, D. H., Shapley, R. M., & Hawken, M. J. (1993). Macaque VI neurons can signal 'illusory' contours. *Nature*, *365*(6446), 550–552.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE international conference on computer vision*.

Von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, *224*(4654), 1260–1262.

Huang, Y., Gornet, J., Dai, S., Yu, Z., Nguyen, T., Tsao, D., & Anandkumar, A. (2020). Neural networks with recurrent generative feedback. In *Neural information processing systems*.

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolazione omogenea. *Rivista Di Psicologia*, *49*(1), 7–30.

Kanizsa, G. (1976). Subjective contours. *Scientific American*, *234*(4), 48–53.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), Article e1003915.

Kim, B., Reif, E., Wattenberg, M., Bengio, S., & Mozer, M. C. (2021). Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, *4*, 251–263.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. arXiv:1412.6980.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. arXiv:1312.6114.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images* (Master's thesis), University of Toronto.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., Patterson, R. D., Howard III, M. A., Friston, K. J., & Griffiths, T. D. (2011). Predictive coding and pitch processing in the auditory cortex. *Journal of Cognitive Neuroscience*, *23*(10), 3084–3094.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, *98*(4), 1907–1911.

Lee, D.-H., Zhang, S., Fischer, A., & Bengio, Y. (2015). Difference target propagation. In *Machine learning and knowledge discovery in databases* (pp. 498–515). Springer.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*, 335–346.

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. In *Advances in neural information processing systems* (pp. 152–164). URL: https://proceedings.neurips.cc/paper/2018/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf.

Lotter, W., Kreiman, G., & Cox, D. (2017). Deep predictive coding networks for video prediction and unsupervised learning. arXiv:1605.08104.

Lotter, W., Kreiman, G., & Cox, D. (2018). A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. arXiv:1805.10734.

Maheswaranathan, N., McIntosh, L., Kastner, D. B., Melander, J., Brezovec, L., Nayebi, A., Wang, J., Ganguli, S., & Baccus, S. A. (2018). Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. *BioRxiv*, Article 340943.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Millidge, B., Tschantz, A., & Buckley, C. L. (2020). Predictive coding approximates backprop along arbitrary computation graphs. arXiv:2006.04182.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251.

Notredame, C.-E., Pins, D., Deneve, S., & Jardri, R. (2014). What visual illusions teach us about schizophrenia. *Frontiers in Integrative Neuroscience*, *8*, 63.

Nour, M. M., & Nour, J. M. (2015). Perception, illusions and Bayesian inference. *Psychopathology*, *48*(4), 217–221.

Pak, A., Ryu, E., Li, C., & Chubykin, A. A. (2020). Top-down feedback controls the cortical representation of illusory contours in mouse primary visual cortex. *Journal of Neuroscience*, *40*(3), 648–660.

Pan, Y., Chen, M., Yin, J., An, X., Zhang, X., Lu, Y., Gong, H., Li, W., & Wang, W. (2012). Equivalent representation of real and illusory contours in macaque V4. *Journal of Neuroscience*, *32*(20), 6760–6770.

Pang, Z., Choksi, B., O'May, C. B., & VanRullen, R. (2020). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. In *NeurIPS 2020 workshop SVRHM*. URL: https://openreview.net/forum?id=l7Gkd1vQBkC.

Parks, T. E. (2001). Rock's cognitive theory of illusory figures: a commentary. *Perception*, *30*(5), 627–631.

Ramachandran, V., Ruskin, D., Cobb, S., Rogers-Ramachandran, D., & Tyler, C. (1994). On the perception of illusory contours. *Vision Research*, *34*(23), 3145–3152.

Raman, R., & Sarkar, S. (2016). Predictive coding: a possible explanation of filling-in at the blind spot. *PLoS One*, *11*(3), Article e0151194.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Schumann, F. (1918). *Psychologische studien*, *Beiträge zur analyse der gesichtswahrnehmungen*. JA Barth.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, *7*, 1792.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Spillmann, L., & Dresp, B. (1995). Phenomena of illusory form: Can we bridge the gap between levels of explanation? *Perception*, *24*(11), 1333–1364.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*(43), 1551.

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., & Liu, Z. (2018). Deep predictive coding network for object recognition. In *International Conference on Machine Learning* (pp. 5266–5275). PMLR.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136–4160.

Whittington, J. C., & Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation*, *29*(5), 1229–1262.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zelano, C., Mohanty, A., & Gottfried, J. A. (2011). Olfactory predictive codes and stimulus templates in piriform cortex. *Neuron*, *72*(1), 178–187.

3.3 Chapter Conclusion

The study suggests the neural network driven by predictive coding dynamics possesses illusory perception, supporting the possibility that the same dynamics strategy, i.e. predictive coding, might be shared between the network and biological brain.

4 Traveling waves in the deep predictive coding network

4.1 Chapter Introduction

From the first two studies, we learned that: (i) cortical traveling waves may reflect the neural mechanisms of predictive coding in the biological brain, with forward waves conveying prediction errors and backward waves transmitting predictions; (ii) the deep neural network which implements predictive coding could perceive illusory contours like humans, meaning such scheme may underlie certain, if not all, brain functions like illusory perception. A natural question may arise: could such a neural network, sharing the same functional architecture and signal dynamics as the human brain, reproduce its oscillations or even traveling waves? If yes, we may be able to obtain supporting evidence for the first study suggesting that oscillations or traveling waves are indeed the neural correlates of predictive coding, and also for the second study supporting the role of predictive coding as the model of human brains.

In the third study, we further update the same predictive neural network as in the second study by adding biologically plausible time delays and constants between network layers in order to generate oscillations. We will first evaluate whether such a model could generate biologically plausible oscillations and traveling waves between layers. Once we get the desired results, in the second step, we could employ such oscillatory neural network to act as a working model of the physical brain and we may use it to test some unclear or controversial physiological or psychological phenomenon by carefully checking the layers' or artificial neuron's activities, which we cannot easily perform in the biological brain.

4.2 Article 3 (In preparation)

Biologically plausible oscillations generated in a predictive deep neural network

4.2.1 Introduction

Predictive coding is a famous and influential theory in the field of neuroscience. It states that instead of passively receiving the external information and forming perception or decision, the brain holds a hierarchical internal modal that actively interacts with the external stimulation. In this hierarchy, each level predicts the activation of the lower level, with the lowest level representing the outside world; at each level the system computes the difference between real stimulation and the predicted one and sends it upward to update the model for better prediction in the future. The scheme provides explanations for a wide range of neurophysiological and psychological phenomena, which leads to the belief that predictive coding may serve as a unifying computational framework for brain functions including sensation, perception, memory, and so on.

An important aspect of predictive coding is to figure out how it is implemented in the physical brain. Specifically, which brain activity could be used to transmit the resulting predictions and error signals within such a framework? It seems that the most significant activities in the brain are oscillations which are ubiquitous in the brain and are thought to be closely related to various brain functions. Importantly, the oscillations can travel between different regions in the brain hierarchy, forming the so-called 'traveling waves'. This property seems to fit the pattern of dynamical message passing in the hierarchical model hypothesized by the predictive coding theory.

Previous work

The study by Alamia and VanRullen (2019) has already investigated the generation and propagating of oscillations in a simplified predictive coding model. They first built a two-layer model which implements predictive coding dynamics as shown in Figure 23A. They defined the prediction error x_L at t timepoint as the difference between the prediction of y_L at $t-\Delta T$ timepoint and the actual activity y_{L-1} at a level below at t timepoint in Equation (9) where L denotes the levels and ΔT indexes the temporal communication delay between them. The prediction y_L can be updated according to Equation (11) which consists of the bottom-up

error signal x_L with a delay ΔT and the difference between its own state and the top-down prediction from the next higher level. The two crucial parameters τ and τ_D refer to the time constant of neuronal integration and decay respectively.

$$x_L(t) = y_{L-1}(t) - y_L(t - \Delta T) \tag{9}$$

$$\frac{dy_L}{dt} = \frac{1}{\tau} \times x_L(t - \Delta T) + \frac{1}{\tau_D} \times (y_{L+1}(t - \Delta T) - y_L(t)) \tag{10}$$

Their simulation results show that the model could generate alpha-band oscillations with biologically plausible time constants and time delay as shown in Figure 23B. Moreover, they performed a parameter exploration (see Figure 23C) and revealed that alpha-band oscillation could emerge robustly in the predictive coding scheme within a wide range of parameters which are biologically plausible.

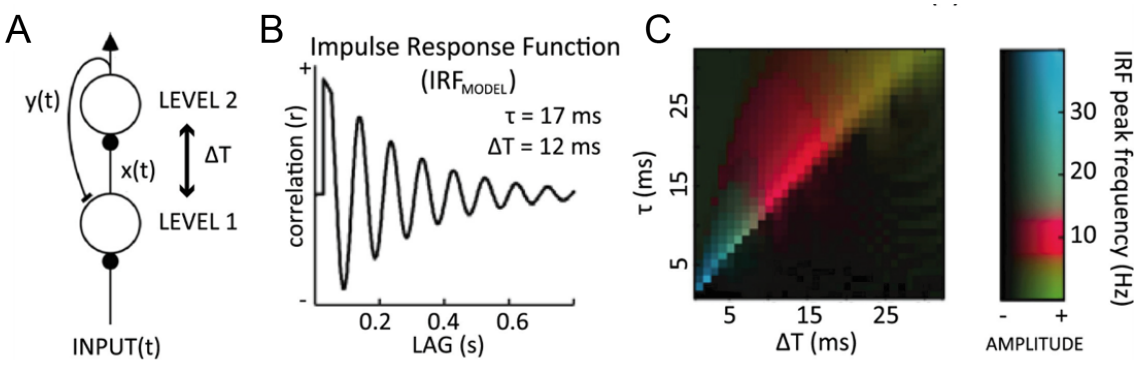


Figure 23: **Alpha oscillations generated in a two-layer predictive coding model.** A. A simple two-layer model implementing predictive coding dynamics. The higher levels make predictions $y(t)$ about the received input by the lower level, and the prediction error $x(t)$ is used to update the next prediction. The time delay between two layers is ΔT. B. the Impulse Response Function (IRF) refers to the results of cross-correlating the input sequence and the corresponding layer's activity, which can extract the responsive pattern of the neuron towards its input. Therefore the figure basically shows the model generating alpha-band oscillations with biologically plausible parameters. C. Systematic exploration of time constant τ and communication delay ΔT suggests that the alpha band oscillations is a robust phenomenon within a biologically plausible range of the values (red colors). Figure adapted from (Alamia and VanRullen, 2019)

In their second experiment, they enlarged the model into a multi-layer architecture to simulate the hierarchical organization of brain regions. Remarkably, the enlarged model could generate forward alpha traveling waves propagating from lower level to higher level with the presence of input data; while opposite waves were dominant when only priors are fed to the highest level. Their results are consistent with empirical observations. All in all, their results suggest the directional alpha traveling waves may underlie the implementation of predictive coding in the brain.

Current study

The work by Alamia and VanRullen (2019) provides crucial evidence of alpha traveling waves as the neural bases of predictive coding. However, their model only has a single function to process univariate data series. By implementing the predictive coding scheme in a deep neural network, we can obtain a versatile model which can deal with multiple tasks, for example, processing visual images or video clips. In the current study, we designed a deep predictive coding neural network according to the algorithm previously devised by Choksi et al. (2021). To meet the current research need, we added time delays and constants between network layers in order to generate oscillations. We expect that such a model could generate biologically plausible oscillations and traveling waves as in Alamia and VanRullen (2019) but also can deal with various cognitive tasks and show similar performance as their human counterpart.

4.2.2 Methods

4.2.2.1 Architecture

We used the same architecture as in Choksi et al. (2021) except additional time delays ΔT between layers were added (see Figure 24). Specifically, the architecture includes N hierarchically stacked autoencoder with N feedforward encoding layers e_n ($n \in [1 \quad N]$) and N corresponding decoding layers d_{n-1}. The feedforward layers perform a convolutional process; while the feedback layers perform a deconvolutional process which attempts to reconstruct the representation at a layer below. The dynamics of information flow inside the model follow a predictive coding strategy: in the hierarchical network, the higher layers try to predict the activity of the lower layers, and the errors made in this prediction are then used to update their activity. The time delays (ΔT) will be considered when information is transmitted between layers including the feedforward input from the lower layer to the higher layer; the feedback signal in the opposite direction; and the error correction signals between two layers.

As shown in Figure 24, the updating of e_n layer needs four parts of information: (i) The feedforward input from the layer below at timepoint $t - \Delta T$ due to the time delay between layers; (ii) The feedback error correction passed from the layer above at timepoint $t - \Delta T$; (iii) The retained memory of the current layer from the last timepoint $t - 1$; and (iv) the feedforward error correction signal. Therefore the update of e_n can be expressed as[1]:

$$e_n(t) = \underbrace{\beta\left[W_{n-1,n}e_{n-1}(t-\Delta T)\right]_+}_{FeedForward} + \underbrace{\lambda\left[W_{n+1,n}e_{n+1}(t-\Delta T)\right]_+}_{FeedBack\ error\ correction} + \underbrace{(1-\beta-\lambda)e_n(t-1)}_{Memory} - \underbrace{\alpha\nabla\epsilon_{n-1}(t-\Delta T)}_{Feedforward\ error\ Correction} \tag{11}$$

where W denotes the weights connecting two adjacent layers with its subscript indicating the

[1]The equation is basically the same as the Equation 1 in Pang et al. (2021), but with communication delays now taken into account

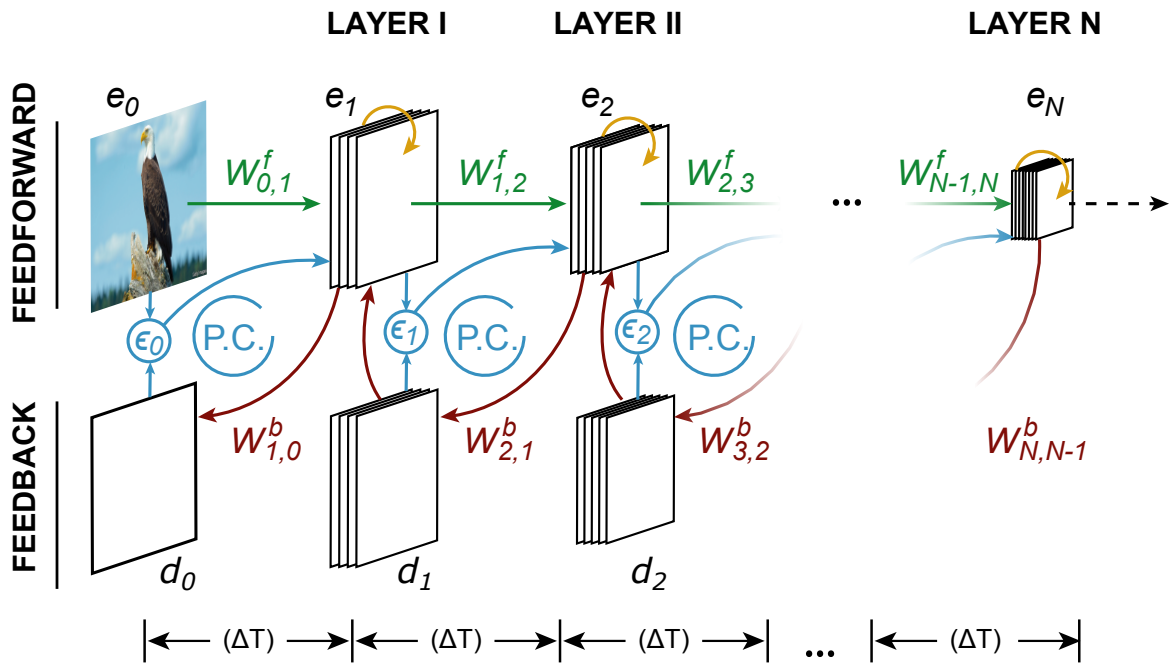


Figure 24: **Network architecture.** The architecture consists of the main body and a possible classification head (depends on the tasks). For the main body, the predictive coding strategy is implemented in stacked autoencoders, with three feedforward encoding layers (e_n) and three generative feedback decoding layers (d_n). Reconstruction errors (ϵ_n) are computed and used for the proposed predictive coding updates which are denoted by 'P.C.' loops. Between the two adjacent layers, a time delay (ΔT) was considered. Figure adapted from (Choksi et al., 2021; Pang et al., 2021)

direction of information flow. For instance, $W_{n-1,n}$ means feedforward weights connecting layer $n-1$ to layer n, and $W_{n+1,n}$ denotes the feedback weights from layer $n+1$ to n. The parameters β, λ and α act as balancing coefficients for the feedforward drive, feedback error correction, and feedforward error correction terms respectively, and they are treated as hyperparameters of the network. Then $\nabla\epsilon_{n-1}(t-\Delta T)$ denotes the gradient of the error at layer $n-1$ with respect to the activation in layer $e_n(t-2\Delta T)$. The reconstruction error for layer $n-1$ can be expressed in Equation (12), which is the mean squared error between the representation $e_{n-1}(t-\Delta T)$ and the corresponding prediction $d_{n-1}(t-\Delta T)$ from layer n

to $n-1$ (see Equation (13)).

$$\epsilon_{n-1}(t-\Delta T) = \|e_{n-1}(t-\Delta T) - d_{n-1}(t-\Delta T)\|_2^2 \tag{12}$$

$$d_{n-1}(t-\Delta T) = W_{n,n-1}e_n(t-2\Delta T) \tag{13}$$

Therefore the feedforward error correction term can be expanded as in Equation (14). After it is replaced in Equation (11), we can get the expanded updating equation for layer e_n as in Equation (15)

$$\begin{aligned}\frac{\partial \epsilon_{n-1}(t-\Delta T)}{\partial e_n(t-2\Delta T)} &= \frac{\partial \|e_{n-1}(t-\Delta T) - W_{n,n-1}e_n(t-2\Delta T)\|_2^2}{\partial e_n(t-2\Delta T)} \\ &= -2W_{n,n-1}(e_{n-1}(t-\Delta T) - W_{n,n-1}e_n(t-2\Delta T))\end{aligned} \tag{14}$$

$$\begin{aligned}e_n(t) = &\underbrace{\beta\big[W_{n-1,n}e_{n-1}(t-\Delta T)\big]_+}_{FeedForward} + \underbrace{\lambda\big[W_{n+1,n}e_{n+1}(t-\Delta T)\big]_+}_{FeedBack\ error\ correction} + \underbrace{(1-\beta-\lambda)e_n(t-1)}_{Memory} \\ &+ \underbrace{2\alpha W_{n,n-1}e_{n-1}(t-\Delta T) - 2\alpha W_{n,n-1}e_n(t-2\Delta T)}_{Feedforward\ error\ Correction}\end{aligned} \tag{15}$$

4.2.2.2 Generation of oscillations

Now we have the architecture and its updating equation for each layer's activities. How do the oscillations come from? For a given layer $e_n(t)$, the oscillations of it depend on its own information at past timepoint. To clearly understand this. Figure 25 shows the information flow which is unfolded through time for updating layer $e_n(t)$. As we can see, the past information for $e_n(t)$ comes from two parts, the instant memory $e_n(t-1)$ at last timepoint as well as $e_n(t-2\Delta T)$. Clearly, the reason for $e_n(t)$ to oscillate is its past information at $t-2\Delta T$. Also because of this when initializing the network for recurrent updating, we need to initialize the feedforward pass and feedback pass for at least $2\Delta T$ timesteps.

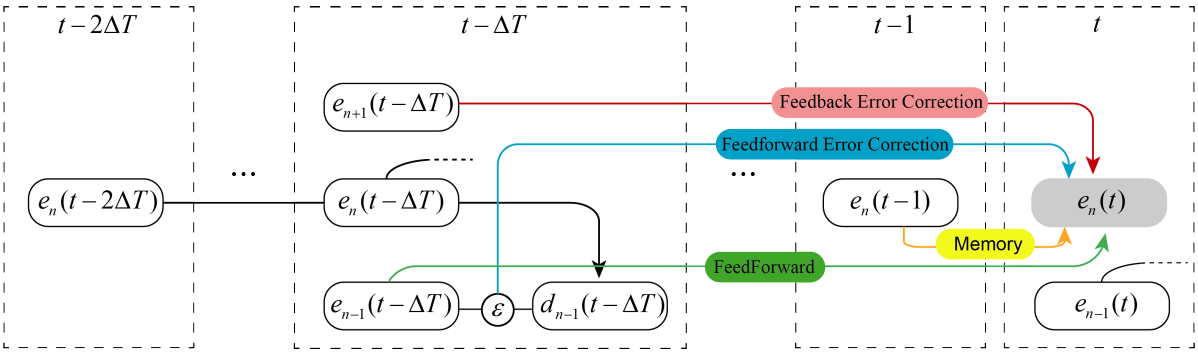


Figure 25: **Unfolded updating process through time.** The updating of layer $e_n(t)$ needs four parts of signals: bottom-up feedforward input at $t-\Delta T$, a memory from the same layer at last timestep, feedback error correction at $t-\Delta T$, and feedforward error correction which involves information at $t-\Delta T$ and $t-2\Delta T$.

Mathematical demonstration of oscillations

Now we have an intuitive understanding of how e_n depends on its past signals to generate oscillations. How can we describe it mathematically? From Equation (15), we can see easily that the 'Memory' term and the second term in the 'Feedforward error Correction' group contain the past information of e_n which will contribute to its oscillatory dynamics. The other terms describing the activity at a level above or below seem irrelevant to e_n's oscillations. However, these terms have a direct connection with the layer of e_n through feedforward or feedback connections as shown in Figure 24. That is, they may contain past information of e_n. Therefore we need to expand all those terms by using the updating Equation (15) for a given layer. But before that, let us first transform the Equation (15) into the form in Equation (16) to express the change of activity for $e_n(t)$.

$$\begin{aligned} e_n(t) - e_n(t-1) = {} & \beta\left[W_{n-1,n}e_{n-1}(t-\Delta T)\right]_+ + \lambda\left[W_{n+1,n}e_{n+1}(t-\Delta T)\right]_+ \\ & - (\beta+\lambda)e_n(t-1) \\ & + 2\alpha W_{n,n-1}e_{n-1}(t-\Delta T) - 2\alpha W_{n,n-1}e_n(t-2\Delta T) \end{aligned} \tag{16}$$

Term expansion. For clarity, I use a graphic demonstration to show how those terms are expanded based on the updating Equation (15). Note again, we expand those terms in order

to reveal the hidden information about e_n at past timepoint. Inside the 'FeedForward' term which is in green, it includes a top-down input from e_n which is marked with a red underline to indicate the 'feedback' connection (check this in Figure 24). Inside the 'FeedBack error correction' term in red, it contains two input streams from lower e_n (check this in Figure 24) which are marked with green and blue underlines. Inside the first part of the 'Feedforward Error Correction' term, it contains a top-down input from e_n marked by a red underline. Next we only need to pick and group those marked terms and ignore other irrelevant terms.

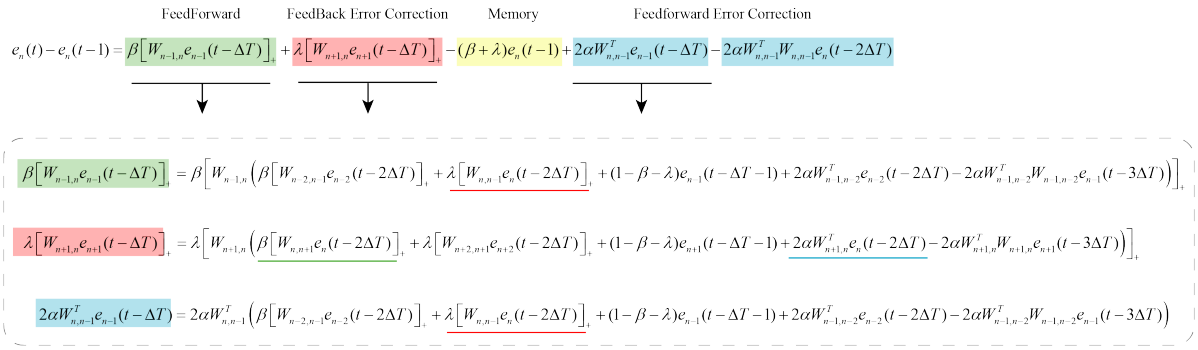


Figure 26: .

Extracting and Grouping. Now we need to pick the related terms as marked in Figure 26 and group them together in order to know how the past information of e_n affects the oscillatory activities. However, from Figure 26, we found that due to the ReLU nonlinearities $[\ ]_+$ as well as the weights parameters W, it's impossible to perform grouping. Therefore, we introduced two simplification hypotheses here: (i) We ignore the ReLU nonlinearities for now; (ii) we set weights as 1×1 kernels with value 1, such that they only copy what they have. Then after grouping all the $e_n(t - 2\Delta T)$ terms and $e_n(t)$ terms, we finally get Equation (17) which shows all the factors that can affect the generation of oscillations for $e_n(t)$. The first term denotes the influence of past information is controlled by the coefficient combination $2\lambda\beta + 4\alpha\lambda - 2\alpha$ and the second decay term is modulated by $\beta + \lambda$. With all those parameters that can produce influence, a critical question is how each parameter affects the oscillatory activities.

$$e_n(t) - e_n(t-1) = (2\lambda\beta + 4\alpha\lambda - 2\alpha)e_n(t - 2\Delta T) - (\beta + \lambda)e_n(t) \tag{17}$$

A generic analytical solution for oscillatory activities

In order to find how each parameter can play a role in the oscillations' pattern. Here we attempt to derive a generic solution for Equation (17). Before we can solve it, a few transformation needs to be carried out. First, to avoid the confusion between layer's activation e_n and the mathematical constant e, I will set $y(t) = e_n(t)$. Second, to solve this differential equation, we need to transform it into a continuous form, i.e. set $\frac{dy}{dt} = e_n(t) - e_n(t-1)$. Third, during simulation, we set one timestep as $0.001s$, thus $dt = 0.001s$. To make both sides of the equation equal, the right side of the equation also need to divided by dt, i.e., multiplied by 1000. Now, the coefficients at right side become $1000 \times (2\lambda\beta + 4\alpha\lambda - 2\alpha)$ and $1000 \times (\beta + \lambda)$ and we can use A and B to denote them for simplified operation. Now Equation (17) becomes:

$$\frac{dy}{dt} = Ay(t - 2\Delta T) - By(t) \tag{18}$$

To solve Equation (18), we can consider a general solution for it as the exponential function in Equation (19). θ can be expressed in Equation (20). where R denotes the real part of the complex number θ, i is the imaginary operator and ω is related to the oscillatory frequency. If θ represents an oscillatory event, meaning the ω is not zero, Then the value of R will influence the oscillatory dynamics in three ways: (i) $R = 0$ means a pure oscillation; (ii) $R > 0$ means the amplitude of the oscillation will become increasingly larger, finally forming an explosion pattern; (iii) $R < 0$, in turn, represents a vanishing oscillation with increasing smaller amplitude.

$$y(t) = e^{\theta t} \tag{19}$$

$$\theta = R + i\omega \tag{20}$$

Based on the general solution in Equation (19), we can compute each side of Equation (18) to get Equation (21) which could be further simplified by removing $e^{\theta t}$ from both sides, as in Equation (22)

$$\theta e^{\theta t} = Ae^{\theta t}e^{-2\theta\Delta T} - Be^{\theta t} \tag{21}$$

$$\theta = Ae^{-2\theta\Delta T} - B \tag{22}$$

During simulation, we will systematically assign values for A, B and ΔT. Thus the only unknown θ can be represented by those known letters. The value of θ computed by the calculator turns out to be a 'Lambert W function' in Equation (23).

$$\theta = \frac{lambertw(0, 2A\Delta Te^{2B\Delta T})}{2\Delta T} - B \tag{23}$$

Clearly, it is difficult to read out the R and ω values from the above solution. Therefore we try another avenue. Equation (23) can be reinjected into Equation (19), from which we can get the oscillatory activities. Such that, we could still use the numerically generated activities from equation to predict the activities from a deep neural network. But before we can use it as a predictor, we first performed a sanity check to prove its validity.

Sanity check for this generic solution

In order to investigate whether our derived generic solution could provide accurate and validate prediction for the model's behaviour, we performed a sanity check by comparing with the results in Alamia and VanRullen (2019). Their model employed a simplified version of predictive coding scheme, thus they don't have feedforward or feedback input to update their model, but only with prediction error information. Accordingly, their analytical solution does not involve the λ and β coefficients. Also, they assumed a pure oscillation, i.e., $\theta = 0 + i\omega$ with $R = 0$. Therefore, the analytical solution can only predict their model's activities only when oscillations are produced, i.e., the color edge part in Figure 23C.

To test our generic solution in their model, we need to further simplify our model by setting both λ and β as zero. However, we did not set $R = 0$, instead we used R as a parameter to predict any activity pattern generated by the model. Therefore the value of A becomes $1000 \times (-2\alpha)$ and B is zero in order to fit the model in Alamia and VanRullen (2019). In

Alamia and VanRullen (2019), after combining Equations (9) and (11), the time constant is $-\frac{1}{\tau}$ which is in millisecond unit, while our unit for simulating is in second. After changing into the same unit, we can get $\alpha = \frac{1}{2\tau}$.

Figure 27A shows the results for the simplified version of our model (i.e., no $beta$ and λ). The curves are the artificial neurons' activity over 1000 timesteps (or ms). The spectrum shows the amplitude of different frequencies under a systematic combination of α and ΔT. As can be seen, there are mainly three types of activity patterns. First, the activities located on the color edge of the spectrum tend to show a steady oscillatory pattern which corresponds to $R = 0$,i.e., a pure oscillation. Second, the activities above the edge show fast vanishing patter, i.e., $R < 0$. Third, the activities below the edge show an explosion pattern meaning $R > 0$. Figure 27B are the spectrum and temporal courses of neuron's activity. Obviously, both results from the model and the equation are almost the same, suggesting that our generic solution computed from the equation could provide a good prediction for the model's performance.

When compared our results with that in Alamia and VanRullen (2019), both show similar parameters' range for the generation of alpha oscillations. As can be seen, the range for the emergence of alpha oscillations is with $\Delta T \in [10\ 20ms]$ and $\alpha \in [0.02\ 0.04]$. Since $\alpha = \frac{1}{2\tau}$, so the range for τ is between 12 and 25ms, which is consistent with previous results as in Figure 23C.

Summary

To summarise, the simplified version of our model showed more generic properties than the model used in Alamia and VanRullen (2019). First, our model considered more parameters that can make a difference. Second, our model adopted a more generic solution to predict all the patterns of the model's behavior including pure oscillation, vanishing, and explosion situations. Alamia and VanRullen (2019) only calculated a specific solution to predict with pure oscillations.

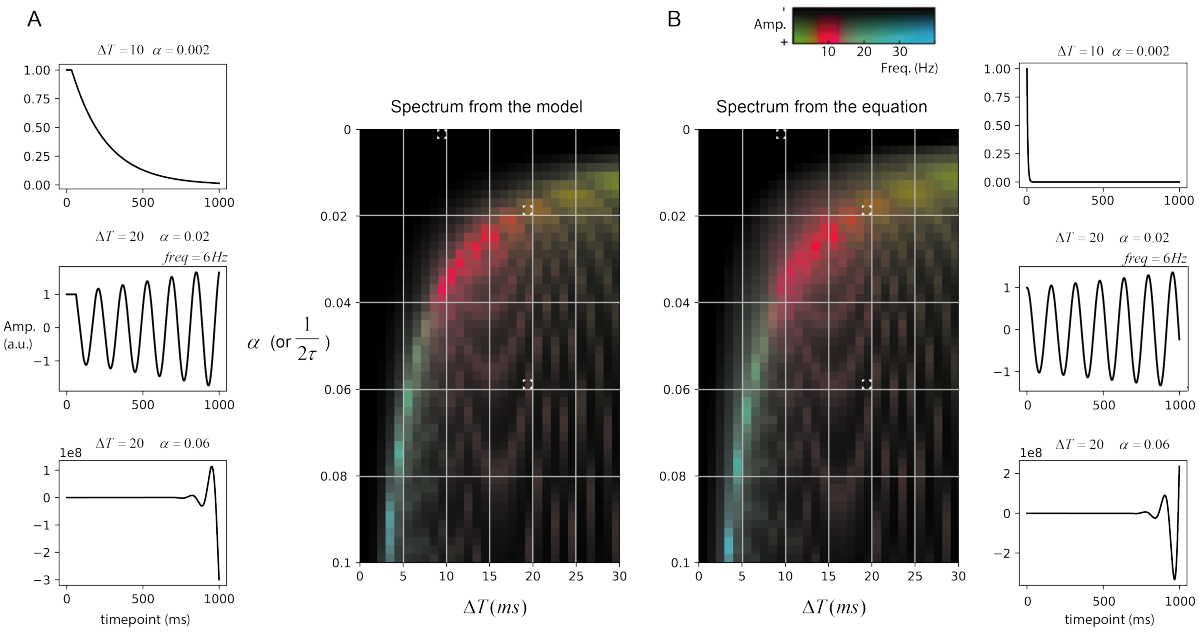


Figure 27: **Oscillations generated in the deep neural network and equation.** A. The simulated neuron's activity and the corresponding spectrum for a systematic combination of parameters. B. The simulated activities from differential equations, which show a similar pattern as the results from the model simulation.

4.2.2.3 Generation of traveling waves

Once our small model (used above) can generate oscillation within biologically plausible parameters' range in terms of α, β, λ and ΔT, the next step is to enlarge the model into multiple layers like what had been done by Alamia and VanRullen (2019). This enlarged model could act as a model of the hierarchical organization of different brain layers. We will examine how the oscillations could be conveyed through this model.

4.2.2.4 Working model of brain function

If the deep neural network could demonstrate that the traveling waves play a critical role in conveying the dynamic messages in the predictive coding framework, we can use it as a working model of the brain. As we know that except for the complexity of the brain itself, due to limitations on brain research methods, we cannot obtain more direct data or evidence about it. However, things will be easier if we have a brain-like model. By carefully examining

this model, we may be able to resolve some debates on the brain.

4.3 Chapter Conclusion

In the third study, we further update the same predictive neural network by adding biologically plausible time delays and constants between network layers in order to generate oscillations. The preliminary results show that the network could oscillate with biologically plausible time parameters. We expect that such an oscillatory neural network will produce more human-like results in terms of its signal unit activation pattern and final decision output

5 General Discussion

The research topic of this thesis is 'Predictive Coding', which is a theory about brain function in the field of neuroscience. Is there a unifying principle that the brain abides by to operate and generate various cognitive functions? The answer given by predictive coding may be 'yes' and under its framework, the simple job of the brain is just to minimize a 'prediction error'. The theory supposes the brain holds a hierarchical inner model of the outside world. When receiving the outside information through the sensory organs (i.e., eyes or ears), our brain constantly predicts what happens at each hierarchical level instead of being a passive feature analyzer. The resulting 'prediction errors' at each level are what the brain tries to minimize and this process is thought to correspond to various functions. For example, correctly identifying a vague sound is considered the result of decreasing the prediction errors by the brain (Blank et al., 2018).

Although predictive coding could serve as a promising theory of the brain, it faces some key issues. First, it may not be possible to test every brain function and examine whether predictive coding could provide a reasonable explanation for them. Therefore, while an increasing amount of works have offered supporting evidence, it does not mean that the brain must follow a predictive coding process. Second, what're the neural bases of the predictive coding in the brain? The predictive coding model supposes a hierarchical structure. How should we fit this hypothesized one to the realistic biological structures in terms of the brain's organization between different regions at a larger scale or between different laminar layers at a smaller scale? Moreover, how does the brain transmit the dynamic exchange of prediction and errors signals proposed by the theory?

5.1 Summary of the thesis

5.1.1 Aim of the thesis

The current thesis tried to solve these questions by evaluating predictive coding theory in both the biological brain and the deep neural networks. On the one hand, we tried to find the

neural dynamics in the brain that relate to the descending prediction signals from a higher level to a lower level as well as the ascending prediction errors signals manifesting when the mismatch between prediction and observed evidence happened. On the other hand, we took advantage of deep neural networks which are supposed to loosely mimic the brain structure and human behaviors such as computer vision. The idea is that if the brain adopts a predictive coding strategy to power its functions, a deep neural network which implements prediction coding dynamics should also present some human-like performance.

Specifically, in the brain, the most predominant activities are brain oscillations. A large body of studies has shown that these oscillations could play a role in information coding (Kayser et al., 2009; Montemurro et al., 2008; O'Keefe and Recce, 1993; Vinck et al., 2010). Particularly, It has been reported that these oscillations could travel between brain regions, forming traveling waves. Could oscillations or traveling waves be related to predictive dynamics? A modeling study by Alamia and VanRullen (2019) suggested that alpha oscillations and traveling waves could naturally appear within a model implementing the simplified vision of predictive coding. Their work provides an insight into the implementation of predictive coding in the biological brain. Could we prove that the traveling waves can relate to the predictive process performed by the brain?

The brain itself is highly complicated, it might not be easy to isolate specific brain structures or neural activities to examine their function without perturbing other neural factors. However deep neural networks could serve as a clean and isolated 'sandbox' where we could only introduce the predictive coding effects and then compare them with the ones generated by brains. Thus the main question here is whether the performance of a predictive deep neural network is similar to that of a physical brain, which could indicate whether the brain employs a predictive coding principle.

Once we get such a model, we could go a bit further in terms of what we may discover in the brain. That is, if we find certain neural activities or neural correlates of predictive coding in the brain, could we reproduce them by virtue of our model which may need to be adapted

for the specific purpose. If yes, then the current thesis could approach the same question from two directions as shown in Figure 28. Particularly, such a scheme could provide mutual corroborating evidence, which is stronger than what we can get from using a single method.

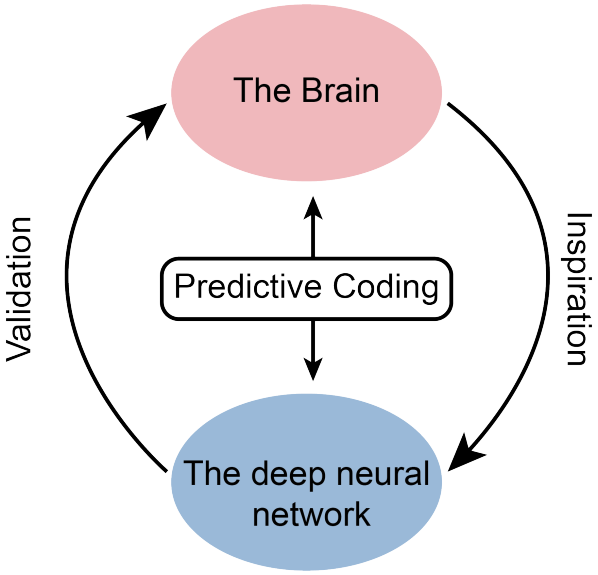


Figure 28: **The methodology of the thesis.** The thesis examined the predictive coding in both the brain and the deep neural network. The deep neural network obtained its inspiration in terms of its construction. If the brain adopts a predictive coding scheme and we could discover its potential underlying neural mechanism or activities in the physical brain, we could build a deep neural network implementing predictive coding and check whether it shows the same activities as in the brain.

5.1.2 Main Results

The first study

The first study examined the nature of traveling waves in the brain. We created two main visual conditions, one with visual input (eyes open) and the other without (eyes closed). We hypothesized that different visual conditions may affect the direction of oscillatory waves traveling across the brain cortex. Our results suggest that the ascending traveling waves only appear with the presence of bottom-up driven visual stimuli and disappear when visual inputs are absent; the descending waves, although they receive some modulation from external visual input, are less affected.

The results could drop a hint at the relationship between traveling waves and predictive coding in terms of how we designed the experiments. When eyes were open, we presented subjects with visual stimuli (static or dynamic discs), which is the situation when both top-down prior expectations and bottom-up driven inputs (or prediction errors) exist. While in the eyes closed situation, only priors were present. If traveling waves could act as the neural mechanism of predictive coding, then we would expect that, during eyes open (or Stimulus-On) period, both forward and backward traveling waves that carry predictions and error signals should exist; while during eyes closed (or Stimulus-Off) period where only priors are presented, we should only expect traveling waves conveying predictions propagate from high-level to low-level regions. This is consistent with what we observed in the experiments. Importantly, the results are in accordance with the modeling study by Alamia and VanRullen (2019), where they showed that oscillatory traveling waves might be a neural signature of predictive coding with forward waves carrying prediction errors and backward waves transmitting prediction signals.

The Second study

In the second study, we build a small deep neural network according to the algorithm previously devised by Choksi et al. (2021) which can implement predictive coding dynamics. We examined whether such a neural network could manifest some human-like performance. Particularly we tested how the model process the Kanizsa shapes from which human observers could perceive the existence of illusory contours. Usual deep neural networks could only recognize the component inducers, i.e., the local features. Could our predictive coding model 'see' the global illusory contours, meaning possessing the illusory perception like humans. The results show that the model reported a higher probability for illusory contours than control conditions.

Some neural networks have been reported to have the ability to perceive illusion which may be because the network could only recognize some particular local features instead of forming illusory contours or edges. In this study, we examined the feedback generative construction of the Kanizsa figure of the model. Compared to the original input figure with no connecting

edge between inducers, the reconstructed image showed the extra illusory edges, indicating the obtained illusory perception by the model. Additionally, the ablation studies demonstrate the direct influence of the prediction-error correction component. Stronger correction for prediction error signals could decrease the probability of perceiving the illusory contours; while weaker or no correction led to stronger illusory perception. This is consistent with the fact that the formation of illusions is due to the deviation of our perception from the physical reality. In summary, through implementing the predictive coding dynamics, the model could obtain some human-like behavior, i.e., illusory perception, which suggests that the brain may also employ such principle.

The Third study

According to the first two studies, we know that first, traveling waves may reflect the neural mechanisms of predictive coding in the biological brain; and second, predictive coding could enable a deep neural network to gain illusory perception. Could such a model generate biologically plausible traveling waves? This question is natural and also critical to further provide supporting evidence in a different direction (from model to brain) to prove predictive coding as a unifying model of brain functions. Therefore in the third study, we further update the same predictive neural network by adding biologically plausible time delays and constants between network layers in order to generate oscillations. The preliminary results show that the network could oscillate with biologically plausible time parameters. We expect that such an oscillatory neural network will produce more human-like results in terms of its signal unit activation pattern and final decision output.

The significance of the third study is reflected in two aspects. First, the existence of oscillations and traveling waves in the predictive coding could provide mutual corroborating evidence for the first study showing the cortical oscillations could convey the prediction and error information in a predictive coding framework. Second, we could use it as a working model of the brain to test and validate the neural activity of some unclear biological phenomena. For example, alpha oscillations have been linked to an inhibitory function (Jensen and Mazaheri, 2010) and other studies suggest an active processing role of alpha oscillations

(VanRullen and Macdonald, 2012). In this case, we may turn to our model and check how each artificial neuron acts when it succeeds or fails to process the corresponding image patch by checking the generative reconstruction image.

5.2 Does the brain function by the principle of predictive coding?

5.2.1 Physical bases for predictive coding in the biological brain

One branch of the thesis is that we explored the neural bases of predictive coding in the brain. In fact, the neural bases could be divided into two aspects: their structural and dynamic bases. The former refers to the computational architectures entailed by predictive coding. As we know, the organization of brain structure can extend across multiple scales ranging from neural populations to large brain regions. So on which spatial scale is predictive coding performed? In terms of dynamic bases, the first and third studies supported the critical role of oscillations and traveling waves. Based on the existing evidence in the literature and the studies in the current thesis, here I argue that the dynamic bases discovered might also indicate the structural bases of predictive coding in the brain.

The most obvious structure for passing the signals is the reciprocal connection between brain regions. Especially, it has been shown that forward connections could convey bottom-up driven information; while backward connections transmit top-down contextual signals which can influence the activities at the lower level. For example, the formation of illusory contours in monkeys is proven to be linked with the feedback signals from V2 and higher regions to V1 (Lee and Nguyen, 2001; Pak et al., 2020; Cox et al., 2013; Pan et al., 2012). When the higher regions are deactivated, meaning no feedback information is available, the illusory perception will disappear (Lee and Nguyen, 2001). Such bidirectional message passing could directly correspond to the top-down predictions and bottom-up errors signals.

Aside from the extrinsic hierarchical connections between layers that could serve as a possible candidate for passing the predictive coding related signals, the intrinsic connectivity of cortical microcircuits may also be possible for routing signals under a predictive coding

framework. For example, many studies have described the possibility of using predictive coding theory to explain the cortical microcircuit process (Bastos et al., 2012; Shipp et al., 2013; Shipp, 2016). As shown in Figure 29, Error signals are supposed to be encoded mainly in layer IV, i.e., the granular layers; while predictions would be encoded in layers II and III. These mechanisms have been identified in many brain areas like primary sensory cortices, motor cortices, parietal cortex, etc (Friston et al., 2017; Owens et al., 2018).

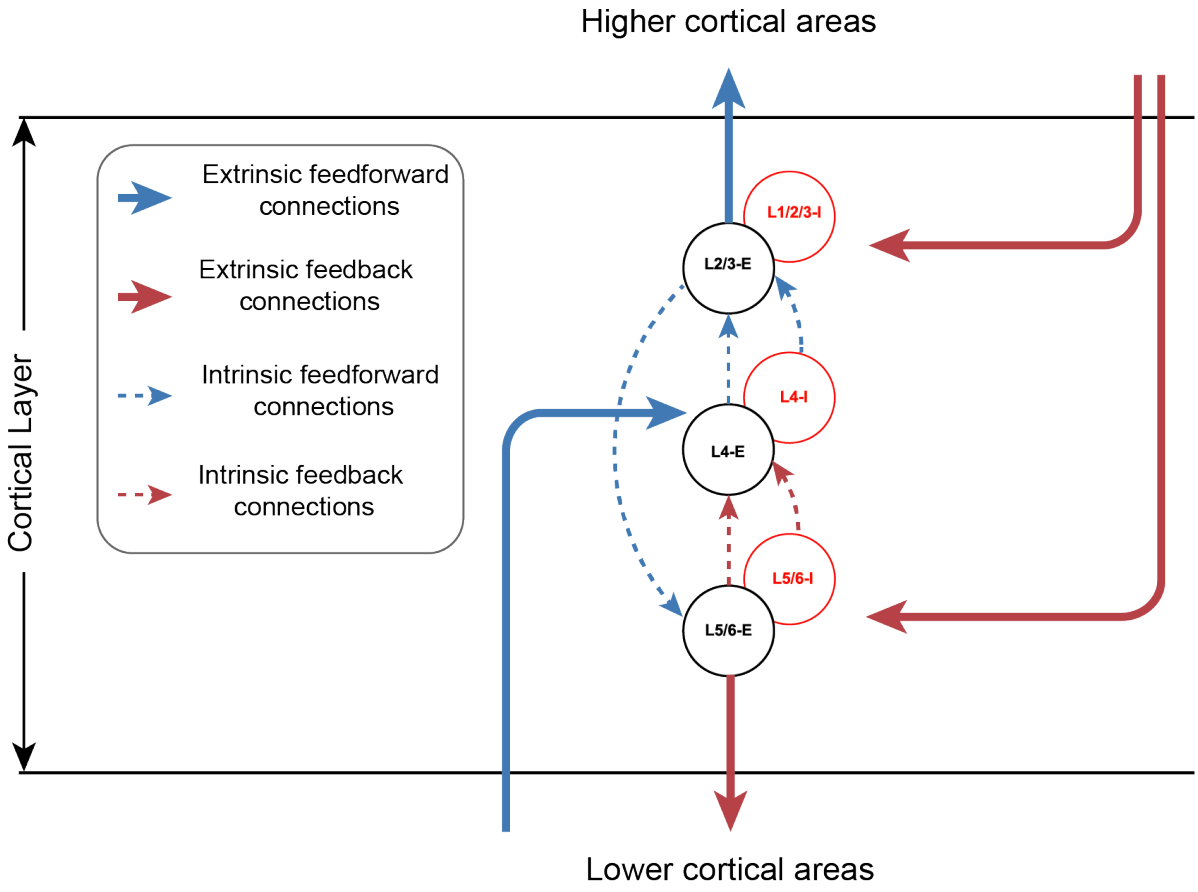


Figure 29: **The canonical cortical microcircuit.** The 6 layers of the cortex could be divided into three-level. Extrinsic connectivity is shown with solid lines and intrinsic connectivity is drawn with dashed lines. Excitatory neural populations are depicted by black circles (E) and Inhibitory ones are red circles (I). Figure adpated from Bastos et al. (2012)

Which pathway connectivity could better match the thrust of the computational strategy of predictive coding? Or, is it possible both extrinsic and intrinsic connections could underlie

the predictive coding process. I will revisit it during the discussion on the dynamic bases of predictive coding in the next part.

5.2.2 Evidence in the current thesis

In the current thesis, we provided both experimental and theoretical evidence to show that the biological brain may adopt a predictive coding strategy to support various brain functions. The first study provides the experimental evidence showing that alpha band waves confirm the PC predictions. The two modeling studies give the theoretical evidence, with the second study simulating the perceptual phenomenon in the pc model, illusory contours and the third study simulating the physiological phenomenon, the oscillatory activities.

5.3 Could brain oscillations be a manifestation of predictive coding signals?

5.3.1 Mutual confirmation between modeling and empirical studies

The thesis adopted a methodology which combined both experimental and modeling research approaches. Both have advantages and disadvantages. For instance, experimentation can provide reliable and direct evidence for the research object. However, it might be time- and resources- consuming. By developing artificial models of human vision we can investigate neuroscientific principles and reduce the need for experimentation. In particular, computational models allow for rapid, iterative experimentation and development, but their results need to be validated by experiment observations.

Combining both methods may provide stronger evidence. In the first study, we performed experiments on human subjects. Although we related the results with predictive coding theory by claiming the traveling waves could underlie the predictive coding dynamics, such explanation is not definite. However, the results from the third modeling study give us more confidence to draw the conclusion that brain oscillations are likely to act as the neural mechanisms of predictive coding dynamics.

Broadly speaking, the benefits of employing the biologically plausible principle in the construction of a deep neural network can be reflected in two aspects. The first one involves the improved performance of a brain-inspired model compared to a standard one. For example, the currently used predictive coding network in the thesis was designed by Choksi et al. (2021) and they pointed out that the introduction of predictive coding dynamics into the popular deep networks like VGG16 and EfficientNetB0 help them improve their robustness against various corruptions. The second aspect concerns how the brain-inspired models, in turn, offer insight into the understanding of the brain in terms of its structure, organization, dynamics and even functions. For instance, the second study suggests that the predictive coding dynamics help the model gain the the illusory perception of contours, from which, we may infer that the two systems (i.e., the deep neural network and the physical brain) with the same perceptual function are likely to share a common underlying mechanism, i.e., the predictive coding.

5.3.2 Consistent findings across different variants of PC model

As we all know, the brain is organized at different scales ranging from the neuron populations to the large brain regions. Based on that, neuronal activities could also be measured at the corresponding level. Therefore, if one wants to build a model of the brain, one question under consideration is which level of abstraction is going to be adopted. In the second and third studies, we used basically similar neural networks. The models can be viewed to model the visual system, with each layer corresponds to a level in the visual hierarchy such as V1 or V2. Each artificial neuron can be viewed as a single neuron or a neuronal population in the biological brain.

In the thesis, the model was used to simulate illusory perception of contours (in the second study) and brain oscillations (in the third study). Here, I argue that the abstraction level adopted by the neural networks are enough for the emergence of those phenomena. In terms of the illusory contour perception, it has been reported that such perception involves the interaction between the activities in V1, V2 and higher visual regions (Lee and Nguyen,

2001; Pak et al., 2020; Cox et al., 2013; Pan et al., 2012). Particularly the feedback path way between those layers are important for the generation of illusion (Lee and Nguyen, 2001). In other words, in order to reproduce this physiological fact, at least two levels of simulation and the reciprocal pathway connections between them are needed structurally. Thus, it is possible to implement in our three-layer model with feedforward and feedback messaging passing between layers.

Brain oscillations can be observed at multiple scales ranging from the activities of individual neurons to that of large neuron populations. However the hypothesized neural mechanism of predictive coding mainly refers to the one related to neuron populations. It has been shown that such large-scale oscillatory activities may result from the interaction between different brain regions, such as the cortical-thalamic circuit (Bollimunta et al., 2011; Contreras et al., 1997). Therefore, we can model the oscillatory activities at the level of neuronal population and ignore the inherent properties of individual neurons. This is what we did for the models used in the second and third studies.

Therefore the abstraction level for the model in the thesis is appropriate for the generation of illusory contours (in the second study) and oscillations (in the third study).

5.3.3 Dynamic bases

The first study demonstrated that forward alpha traveling waves were related to ascending prediction errors and backward waves were linked with descending prediction messages. This is important since we separated the single and mixed wave signals using our wave quantification method. Many researchers have pointed out that it is impossible to observe separate prediction and error signals directly, instead, signals are mixed together (Ficco et al., 2021; Friston, 2005). Therefore, when linking the predictive process to the brain activities, researchers like Friston (2005) suggested that the relevant cortical activities, like the ERPs component P300 can only represent the dynamical exchange between prediction and error signals instead of directly linking to either one. The results of the first study went a bit further by showing the segregation of mixed waves signals into bidirectional message passing

in the predictive coding framework.

Moreover, the results in the first study also pointed out that predictions are related to the spontaneous brain activities since they are present even without the visual stimulation; while the error signals are correlated with evoked responses since they showed up after the presence of visual stimulation and disappeared immediately after the stimulation was cut off. Given the spontaneous response is more related to the entire brain state while the evoked one is more driven by bottom-up stimulation, the results are consistent with the claim by Friston stating that predictions are modulatory and error signals are driving (Friston, 2005). However, in the current literature, most studies only focused on the evoked response in the brain and study how it could be related to the predictive coding dynamics. Later work could pay more attention to spontaneous responses and their potential role in conveying prediction signals.

In the first and third studies, we presented the crucial roles of alpha oscillations and traveling waves in the predictive coding framework. However, as I reviewed in Section 1.5.4., many researchers hold the idea that fast gamma bands could convey forward prediction errors; while slow alpha oscillations could carry backward prediction signals. This seems to contradict the results in this thesis. However, if we postulate that the brain can operate predictive coding at both the large scale i.e. through the interarea reciprocal connections as well as at the local small scale, i.e., through the canonical microcircuit, the two seemingly contradictory transmitting modes can reconcile. That is, the canonical microcircuits rely on ascending fast gamma and beta bands as well as descending alpha oscillations to execute the exchange of predictions and error signals; while at a larger scale, the reciprocal connections can employ both forward and backward alpha traveling waves. This hypothesis is considered based on the fact that fast frequencies such as gamma and beta bands can distribute and generate in relatively smaller regions while slow frequencies like alpha can expand large areas as reviewed in Section 1.5.1.3. Notably, that could also be supported by the third study where we used ΔT to control the time communication delay between adjacent layers. Therefore, smaller ΔT represent spatially close distance and vice versa for a larger value. One can see that

clearly in Figure 27. When ΔT is smaller, then the same model can predict emergence of beta or gamma oscillations: with ΔT around 6-8ms one can get beta and with ΔT around 3-4ms one gets gamma frequency. Now another question to consider is whether the alpha oscillations involved in both the large scale signal exchange and the one at small scale could provide feedback information for both feedforward error signals. Future work is needed to reveal how these signals interact with each other.

5.4 Properties of prediction error signals

From the work in the current thesis, we can extracted some properties for prediction errors and prediction signals respectively, which can be considered together with various empirical evidence. First, we found that predictive error signals can be caused and modulated by visual processing. Second, predictive error signals can correct the perception towards outside world to ground truth.

As for error signals, we found they can be caused and modulated by visual processing. the first study show that FW waves are only present during stimulus-on period. importantly, FW waves have differential pattern under different stimulus types. they are larger for more complex visual stimulation. Secondly, the PE signals can correct the perception to ground truth which can be proved by the results in the second study. In the predictive coding model, when the feedforward error correction was removed, we found that higher illusion was observed, showing that that PE signals could suppress the effect from top-down prior.

5.5 Properties of prediction signals

What's the properties of top-down prediction signals. First, prediction signals originate from endogenous activity. In the first study, we relate BW waves to prediction signals and spontaneous oscillations were closely related to BW waves. Therefore it may be possible that prediction signals can be carried and distributed by endogenous brain activity. Second, prediction signals drive perception to high level prior. Our modeling study in the second work showed that the removal of feedback prediction signal completely eliminates the illusory per-

ception of the network. It demonstrated that the illusory perception is caused the top-down signals in the brain. Third, prediction signals can be modulated by predictive error signals. This is supported by both our first empirical study and second modeling study. In the first study, the presence of visual stimulation reduces BW waves meaning that when top-down prediction can explain bottom-up error signals, part of top-down signals can be cancel out by error signals. In the second study, the removal of error signals causes higher illusory perception of the model, meaning bottom-up signals can suppress and correct the top-down signals.

In summary, both PE signals and prediction signals have their own characteristics and functions, but they can also affect and modulate each other to support the resultant cognitive function operated by the brain.

5.6 Future work

According to the logic of the current thesis, there could also be two branches for future work. First, we can continue the research of predictive coding in the brain and provide more empirical evidence. For example, in the current thesis we considered that alpha band oscillation can serve as the prediction error signals. However, other studies reported that gamma-band oscillation could also play the role. Moreover, many evidence has shown the coupling relationship between gamma and alpha band activities, meaning the coordination between them. Therefore it might be possible that both gamma and alpha activities could serve as PE signals but at different scales. It maybe possible that gamma band oscillation may be critical in local predictive coding circuits in the brain, whereas alpha band in global circuits.

Another branch is to further explore predictive coding in the deep neural networks. On the one hand, the current models in the thesis could be used to validate various phenomena that may be caused by predictive coding. The model in the second study could be used to examine the impact of specific signals for perceptual phenomena such as binding problem.

The benefit of this approach is easy to control and less harm on animals. Also the model in the third study could be used to build direct link between input and resultant oscillatory activities. On the other hand, we can also update predictive coding into other variants like adding predictive coding loops in spiking neural networks.

5.7 Conclusion

The thesis evaluated the predictive coding theory in both the biological brain and the deep neural network. In the brain, our work suggested that the cortical oscillations, especially traveling waves which reflect the oscillations' propagation across the spatial regions in the brain hierarchy, may serve as the potential neural mechanisms of the realization of predictive coding scheme in the brain. In the deep neural networks, the introduction of predictive coding dynamics endows the model with human-like illusory perception of contours, which suggests the possibility that the biological brain may also operate under such a framework. Moreover, aside from the gained illusory perception, when the model was updated to include time communication delays between model layers, it could generate oscillations within biologically plausible time parameters. That is, the likely neural mechanisms (i.e., brain oscillations) of predictive coding emerged naturally from a human-like deep neural network implementing a predictive coding strategy, which further demonstrates the close link between brain oscillations and predictive coding.

In summary, the thesis provides supporting evidence for the notion that predictive coding could act as a unifying framework of brain functions by showing (i) predictive coding dynamics help a deep neural network gain human-like perception; (ii) brain oscillations may offer neural mechanisms for the realization of predictive coding in biological brains.

References

E. D. Adrian and Y. Zotterman. The impulses produced by sensory nerve-endings: Part ii. the response of a single end-organ. The Journal of physiology, 61(2):151–171, 1926.

A. Alamia and R. VanRullen. Alpha oscillations and traveling waves: Signatures of predictive coding? PLoS Biology, 17(10):e3000487, 2019.

L. H. Arnal and A.-L. Giraud. Cortical oscillations and sensory predictions. Trends in cognitive sciences, 16(7):390–398, 2012.

F. Attneave. Some informational aspects of visual perception. Psychological review, 61(3): 183, 1954.

B. Baars. A Cognitive Theory of Consciousness. Cambridge University Press, 1993.

A. Bahramisharif, M. A. van Gerven, E. J. Aarnoutse, M. R. Mercier, T. H. Schwartz, J. J. Foxe, N. F. Ramsey, and O. Jensen. Propagating neocortical gamma bursts are coordinated by traveling alpha waves. Journal of Neuroscience, 33(48):18849–18854, 2013.

D. Baldauf and R. Desimone. Neural mechanisms of object-based attention. Science, 344 (6182):424–427, 2014.

A. Banino, C. Barry, B. Uria, C. Blundell, T. Lillicrap, P. Mirowski, A. Pritzel, M. J. Chadwick, T. Degris, J. Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. Nature, 557(7705):429–433, 2018.

H. B. Barlow. The coding of sensory messages. Current Problems in Animal Behaviour, (331–360), 1961.

A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston. Canonical microcircuits for predictive coding. Neuron, 76(4):695–711, 2012.

A. M. Bastos, V. Litvak, R. Moran, C. A. Bosman, P. Fries, and K. J. Friston. A dcm study of spectral asymmetries in feedforward and feedback connections between visual areas v1 and v4 in the monkey. Neuroimage, 108:460–475, 2015a.

A. M. Bastos, J. Vezoli, C. A. Bosman, J.-M. Schoffelen, R. Oostenveld, J. R. Dowdall, P. De Weerd, H. Kennedy, and P. Fries. Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron, 85(2):390–401, 2015b.

T. J. Baumgarten, A. Schnitzler, and J. Lange. Beta oscillations define discrete perceptual cycles in the somatosensory domain. Proceedings of the National Academy of Sciences, 112(39):12187–12192, 2015.

R. Becker, P. Ritter, and A. Villringer. Influence of ongoing alpha rhythm on the visual evoked potential. Neuroimage, 39(2):707–716, 2008.

H. Berger. Über das elektrenkephalogramm des menschen. Archiv für Psychiatrie und Nervenkrankheiten, 87:527–570, 1929.

G. H. Bishop. Cyclic changes in excitability of the optic pathway of the rabbit. American Journal of Physiology-Legacy Content, 103(1):213–224, 1932.

H. Blank, M. Spangenberg, and M. H. Davis. Neural prediction errors distinguish perception and misperception of speech. Journal of Neuroscience, 38(27):6076–6089, 2018.

A. Bollimunta, J. Mo, C. E. Schroeder, and M. Ding. Neuronal mechanisms and attentional modulation of corticothalamic alpha oscillations. Journal of Neuroscience, 31(13):4935–4943, 2011.

M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis. Reinforcement learning, fast and slow. Trends in cognitive sciences, 23(5):408–422, 2019.

V. Boutin. Etude d'un algorithme hiérarchique de codage épars et prédictif: vers un modèle bio-inspiré de la perception visuelle. PhD thesis, Aix-Marseille, 2020.

V. Boutin, A. Franciosini, F. Chavane, F. Ruffier, and L. Perrinet. Sparse deep predictive coding captures contour integration capabilities of the early visual system. PLoS computational biology, 17(1):e1008629, 2021.

F. Bre, J. M. Gimenez, and V. D. Fachinotti. Prediction of wind pressure coefficients on building surfaces using artificial neural networks. Energy and Buildings, 158:1429–1441, 2018.

T. J. Buschman and E. K. Miller. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. science, 315(5820):1860–1862, 2007.

G. Buzsáki and A. Draguhn. Neuronal oscillations in cortical networks. science, 304(5679): 1926–1929, 2004.

B. Choksi, M. Mozafari, C. B. O'May, B. Ador, A. Alamia, and R. VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. arXiv preprint arXiv:2106.02749, 2021.

A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and brain sciences, 36(3):181–204, 2013.

L. Cohen, S. Dehaene, L. Naccache, S. Lehéricy, G. Dehaene-Lambertz, M.-A. Hénaff, and F. Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. Brain, 123(2):291–307, 2000.

M. X. Cohen and T. H. Donner. Midfrontal conflict-related theta-band power reflects neural oscillations that predict behavior. Journal of neurophysiology, 110(12):2752–2763, 2013.

D. Contreras, A. Destexhe, T. J. Sejnowski, and M. Steriade. Spatiotemporal patterns of spindle oscillations in cortex and thalamus. Journal of Neuroscience, 17(3):1179–1196, 1997.

M. A. Cox, M. C. Schmid, A. J. Peters, R. C. Saunders, D. A. Leopold, and A. Maier. Receptive field focus of visual area v4 neurons determines responses to illusory surfaces. Proceedings of the National Academy of Sciences, 110(42):17095–17100, 2013.

Y. Dan, J. J. Atick, and R. C. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. Journal of neuroscience, 16(10): 3351–3362, 1996.

S. V. David, B. Y. Hayden, and J. L. Gallant. Spectral receptive field properties explain shape selectivity in area v4. Journal of neurophysiology, 96(6):3492–3505, 2006.

P. Dayan and L. Abbott. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Massachusetts Institute of Technology Press, 2001.

S. Dehaene and J.-P. Changeux. Experimental and theoretical approaches to conscious processing. Neuron, 70(2):200–227, 2011.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

L. Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine, 29(6):141–142, 2012.

R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. Annual review of neuroscience, 18(1):193–222, 1995.

R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. Journal of Neuroscience, 4(8):2051–2062, 1984.

R. J. Douglas and K. A. Martin. Behavioral architecture of the cortical sheet. Current Biology, 22(24):R1033–R1038, 2012.

L. Dugué, D. McLelland, M. Lajous, and R. VanRullen. Attention searches nonuniformly in space and in time. Proceedings of the National Academy of Sciences, 112(49):15214–15219, 2015.

M. Eimer. The face-sensitive n170 component of the event-related brain potential. The Oxford handbook of face perception, 28:329–44, 2011.

A. K. Engel and W. Singer. Temporal binding and the neural correlates of sensory awareness. Trends in cognitive sciences, 5(1):16–25, 2001.

A. K. Engel, P. Fries, and W. Singer. Dynamic predictions: oscillations and synchrony in top–down processing. Nature Reviews Neuroscience, 2(10):704–716, 2001.

R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. Nature, 392(6676):598–601, 1998.

G. B. Ermentrout and D. Kleinfeld. Traveling electrical waves in cortex: insights from phase dynamics and speculation on a computational role. Neuron, 29(1):33–44, 2001.

M. Eyesenck and M. Keane. Cognitive Psychology: A Student's Handbook. Psychology Press, 2010.

D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991), 1(1):1–47, 1991.

R. Fellinger, W. Gruber, A. Zauner, R. Freunberger, and W. Klimesch. Evoked traveling alpha waves predict visual-semantic categorization-speed. NeuroImage, 59(4):3379–3388, 2012.

L. Ficco, L. Mancuso, J. Manuello, A. Teneggi, D. Liloia, S. Duca, T. Costa, G. Z. Kovacs, and F. Cauda. Disentangling predictive processing in the brain: A meta-analytic study in favour of a predictive network. 2021.

I. C. Fiebelkorn, Y. B. Saalmann, and S. Kastner. Rhythmic sampling within and between objects despite sustained attention at a cued location. Current Biology, 23(24):2553–2558, 2013.

J. J. Foxe and A. C. Snyder. The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. Frontiers in psychology, 2:154, 2011.

P. Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends in cognitive sciences, 9(10):474–480, 2005.

K. Friston. A theory of cortical responses. Philosophical transactions of the Royal Society B: Biological sciences, 360(1456):815–836, 2005.

K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo. Active inference: a process theory. Neural computation, 29(1):1–49, 2017.

K. J. Friston. Waves of prediction. PLoS biology, 17(10):e3000426, 2019.

K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and cooperation in neural nets, pages 267–285. Springer, 1982.

R. Gaillard, S. Dehaene, C. Adam, S. Clémenceau, D. Hasboun, M. Baulac, L. Cohen, and L. Naccache. Converging intracranial markers of conscious access. PLoS biology, 7(3): e1000061, 2009.

W. Gerstner. Spiking neuron models: single neurons, populations, plastivity. Cambridge University Press, 2002.

C. G. Gross, C. d. Rocha-Miranda, and D. Bender. Visual properties of neurons in inferotemporal cortex of the macaque. Journal of neurophysiology, 35(1):96–111, 1972.

R. Gulbinaite, H. van Rijn, and M. X. Cohen. Fronto-parietal network oscillations reveal relationship between working memory capacity and cognitive control. Frontiers in human neuroscience, 8:761, 2014.

H. Haken. Principles of brain functioning. Springer, 1996.

M. Halgren, I. Ulbert, H. Bastuji, D. Fabó, L. Erőss, M. Rey, O. Devinsky, W. K. Doyle, R. Mak-McCully, E. Halgren, et al. The generation and propagation of the human alpha rhythm. Proceedings of the National Academy of Sciences, 116(47):23772–23782, 2019.

K. Han, H. Wen, Y. Zhang, D. Fu, E. Culurciello, and Z. Liu. Deep predictive coding network with local recurrent processing for object recognition. arXiv preprint arXiv:1805.07526, 2018.

S. Han and X. He. Modulation of neural activities by enhanced local selection in the processing of compound stimuli. Human Brain Mapping, 19(4):273–281, 2003.

C. Harrison. Experiments with linear prediction in television. Bell System Technical Journal, 31(4):764–783, 1952.

S. Haykin. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.

S. Higgs. Biological psychology. Los Angeles, 2014.

C. C. Hilgetag and A. Goulas. 'hierarchy'in the organization of brain networks. Philosophical Transactions of the Royal Society B, 375(1796):20190319, 2020.

S. A. Hillyard, W. A. Teder-Sälejärvi, and T. F. Münte. Temporal dynamics of early perceptual processing. Current opinion in neurobiology, 8(2):202–210, 1998.

J. Hohwy, A. Roepstorff, and K. Friston. Predictive coding explains binocular rivalry: An epistemological review. Cognition, 108(3):687–701, 2008.

A. O. Holcombe and W.-Y. Chen. Splitting attention reduces temporal resolution from 7 hz for tracking one object to¡ 3 hz when tracking three. Journal of vision, 13(1):12–12, 2013.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology, 148(3):574–591, 1959.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1):106–154, 1962.

Y. Iwamura. Hierarchical somatosensory processing. Current opinion in neurobiology, 8(4): 522–528, 1998.

O. Jensen and A. Mazaheri. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Frontiers in human neuroscience, 4:186, 2010.

P. Johnston, J. Robinson, A. Kokkinakis, S. Ridgeway, M. Simpson, S. Johnson, J. Kaufman, and A. W. Young. Temporal and spatial localization of prediction-error signals in the visual brain. Biological psychology, 125:45–57, 2017.

J. H. Kaas and T. A. Hackett. Subdivisions of auditorycortex and levels of processing in primates. Audiology and Neurotology, 3(2-3):73–85, 1998.

E. Kandel, J. Schwartz, and T. Jessel. Principles of Neural Science. Elsevier, 1991.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience, 17(11):4302–4311, 1997.

C. Kayser, M. A. Montemurro, N. K. Logothetis, and S. Panzeri. Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. Neuron, 61(4): 597–608, 2009.

C. Kiley and W. Usrey. Cortical Processing of Visual Signals. Springer, New York, NY, 2013.

W. Klimesch, S. Hanslmayr, P. Sauseng, W. R. Gruber, and M. Doppelmayr. P1 and traveling alpha waves: evidence for evoked oscillations. Journal of neurophysiology, 97(2):1311–1318, 2007.

N. Kriegeskorte and P. K. Douglas. Cognitive computational neuroscience. Nature neuroscience, 21(9):1148–1160, 2018.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.

S. W. Kuffler. Discharge patterns and functional organization of mammalian retina. Journal of neurophysiology, 16(1):37–68, 1953.

A. N. Landau and P. Fries. Attention samples stimuli rhythmically. Current biology, 22(11): 1000–1004, 2012.

A. N. Landau, H. M. Schreyer, S. Van Pelt, and P. Fries. Distributed attention is implemented through theta-rhythmic gamma modulation. Current Biology, 25(17):2332–2337, 2015.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

T. S. Lee and M. Nguyen. Dynamics of subjective contour formation in the early visual cortex. Proceedings of the National Academy of Sciences, 98(4):1907–1911, 2001.

L. Levita, P. Howsley, J. Jordan, and P. Johnston. Potentiation of the early visual response to learned danger signals in adults and adolescents. Social cognitive and affective neuroscience, 10(2):269–277, 2015.

J. E. Lisman and M. A. Idiart. Storage of 7+/-2 short-term memories in oscillatory subcycles. Science, 267(5203):1512–1515, 1995.

J. Liu, M. Higuchi, A. Marantz, and N. Kanwisher. The selectivity of the occipitotemporal m170 for faces. Neuroreport, 11(2):337–341, 2000.

M. Livingstone, D. Hubel, et al. Segregation of depth: form, anatomy, color, physiology, and movement, and perception. Science, 240(4853):740–749, 1988.

W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104, 2016.

D. Lozano-Soldevilla and R. VanRullen. The hidden spatial dimension of alpha: 10-hz perceptual echoes propagate as periodic traveling waves in the human brain. Cell reports, 26 (2):374–380, 2019.

S. Luck and E. Kappenman. The Oxford Handbook of Event-Related Potential Components. Oxford University Press, 2012.

S. J. Luck. An introduction to the event-related potential technique. MIT press, 2014.

S. J. Luck, G. F. Woodman, and E. K. Vogel. Event-related potential studies of attention. Trends in cognitive sciences, 4(11):432–440, 2000.

S. Makeig, M. Westerfield, T.-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. J. Sejnowski. Dynamic brain sources of visual evoked responses. Science, 295(5555):690–694, 2002.

J. Makhoul. Linear prediction: A tutorial review. Proceedings of the IEEE, 63(4):561–580, 1975.

A. Mazaheri and O. Jensen. Rhythmic pulsing: linking ongoing brain activity with evoked responses. Frontiers in human neuroscience, 4:177, 2010.

B. D. McCandliss, L. Cohen, and S. Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. Trends in cognitive sciences, 7(7):293–299, 2003.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133, 1943.

L. Melloni, C. Molina, M. Pena, D. Torres, W. Singer, and E. Rodriguez. Synchronization of neural activity across cortical areas correlates with conscious perception. Journal of neuroscience, 27(11):2858–2865, 2007.

M.-M. Mesulam. From sensation to cognition. Brain: a journal of neurology, 121(6):1013–1052, 1998.

D. Meunier, R. Lambiotte, A. Fornito, K. Ersche, and E. T. Bullmore. Hierarchical modularity in human brain functional networks. Frontiers in neuroinformatics, 3:37, 2009.

G. Michalareas, J. Vezoli, S. Van Pelt, J.-M. Schoffelen, H. Kennedy, and P. Fries. Alpha-beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical areas. Neuron, 89(2):384–397, 2016.

B. Millidge, A. Seth, and C. L. Buckley. Predictive coding: a theoretical and experimental review. arXiv preprint arXiv:2107.12979, 2021.

M. Mishkin and L. G. Ungerleider. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. Behavioural brain research, 6(1):57–77, 1982.

M. Mishkin, L. G. Ungerleider, and K. A. Macko. Object vision and spatial vision: two cortical pathways. Trends in neurosciences, 6:414–417, 1983.

K. E. Misulis and T. Fakhoury. Spehlmann's Evoked Potential Primer. Butterworth-heinemann, 2011.

M. A. Montemurro, M. J. Rasch, Y. Murayama, N. K. Logothetis, and S. Panzeri. Phase-of-firing coding of natural visual stimuli in primary visual cortex. Current biology, 18(5): 375–380, 2008.

J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. Science, 229(4715):782–784, 1985.

V. B. Mountcastle. The columnar organization of the neocortex. Brain: a journal of neurology, 120(4):701–722, 1997.

L. Muller, F. Chavane, J. Reynolds, and T. J. Sejnowski. Cortical travelling waves: mechanisms and computational principles. Nature Reviews Neuroscience, 19(5):255–268, 2018.

I. Nauhaus, L. Busse, D. L. Ringach, and M. Carandini. Robustness of traveling waves in ongoing activity of visual cortex. Journal of Neuroscience, 32(9):3088–3094, 2012.

C. Neuper and G. Pfurtscheller. Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. International journal of psychophysiology, 43 (1):41–58, 2001.

P. Nunez and R. Srinivasan. Electric Fields of the Brain: The Neurophysics of EEG. Oxford University Press, 2006.

J. O'Keefe and M. L. Recce. Phase relationship between hippocampal place units and the eeg theta rhythm. Hippocampus, 3(3):317–330, 1993.

D. O'Shaughnessy. Linear predictive coding. IEEE potentials, 7(1):29–32, 1988.

A. P. Owens, M. Allen, S. Ondobaka, and K. J. Friston. Interoceptive inference: from computational neuroscience to clinic. Neuroscience & Biobehavioral Reviews, 90:174–183, 2018.

A. Pak, E. Ryu, C. Li, and A. A. Chubykin. Top-down feedback controls the cortical representation of illusory contours in mouse primary visual cortex. Journal of Neuroscience, 40 (3):648–660, 2020.

Y. Pan, M. Chen, J. Yin, X. An, X. Zhang, Y. Lu, H. Gong, W. Li, and W. Wang. Equivalent representation of real and illusory contours in macaque v4. Journal of Neuroscience, 32 (20):6760–6770, 2012.

Z. Pang, C. B. O'May, B. Choksi, and R. VanRullen. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. arXiv preprint arXiv:2102.01955, 2021.

A. Pasupathy and C. E. Connor. Population coding of shape in area v4. Nature neuroscience, 5(12):1332–1338, 2002.

T. M. Patten, C. J. Rennie, P. A. Robinson, and P. Gong. Human cortical traveling waves: dynamical properties and correlations with responses. PLoS One, 7(6):e38392, 2012.

M. Penttonen and G. Buzsáki. Natural logarithmic relationship between brain oscillators. Thalamus & Related Systems, 2(2):145–152, 2003.

M. I. Posner and S. E. Petersen. The attention system of the human brain. Annual review of neuroscience, 13(1):25–42, 1990.

R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience, 2(1):79–87, 1999.

B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, et al. A deep learning framework for neuroscience. Nature neuroscience, 22(11):1761–1770, 2019.

J. E. Robinson, M. Breakspear, A. W. Young, and P. J. Johnston. Dose-dependent modulation of the visually evoked n1/n170 by perceptual surprise: a clear demonstration of prediction-error signalling. European Journal of Neuroscience, 52(11):4442–4452, 2020.

K. S. Rockland and D. N. Pandya. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain research, 179(1):3–20, 1979.

A. K. Roopun, M. A. Kramer, L. M. Carracedo, M. Kaiser, C. H. Davies, R. D. Traub, N. J. Kopell, and M. A. Whittington. Temporal interactions between cortical rhythms. Frontiers in neuroscience, 2:34, 2008.

F. Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

B. Rossion and C. Jacques. Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? ten lessons on the n170. Neuroimage, 39(4):1959–1979, 2008.

B. Rossion, C. A. Joyce, G. W. Cottrell, and M. J. Tarr. Early lateralization and orientation tuning for face, word, and object processing in the visual cortex. Neuroimage, 20(3): 1609–1624, 2003.

D. E. Rumelhart, J. L. McClelland, P. R. Group, et al. Parallel distributed processing, volume 1. IEEE Massachusetts, 1988.

A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. arXiv preprint arXiv:1706.01427, 2017.

T. K. Sato, I. Nauhaus, and M. Carandini. Traveling waves in visual cortex. Neuron, 75(2): 218–229, 2012.

P. Sauseng and W. Klimesch. What does phase information of oscillatory brain activity tell us about cognitive processes? Neuroscience & Biobehavioral Reviews, 32(5):1001–1013, 2008.

M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. Journal of neuroscience, 18(10): 3870–3896, 1998.

C. Shain, I. A. Blank, M. van Schijndel, W. Schuler, and E. Fedorenko. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. Neuropsychologia, 138:107307, 2020.

C. E. Shannon. A mathematical theory of communication. The Bell system technical journal, 27(3):379–423, 1948.

C. E. Shannon and W. Weaver. The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1949.

S. Shipp. Neural elements for predictive coding. Frontiers in psychology, 7:1792, 2016.

S. Shipp, R. A. Adams, and K. J. Friston. Reflections on agranular architecture: predictive coding in the motor cortex. Trends in neurosciences, 36(12):706–716, 2013.

H. A. Simon. The architecture of complexity. Proc. Am. Philos. Soc., 106:467–482, 1962.

M. W. Spratling. Predictive coding as a model of biased competition in visual attention. Vision research, 48(12):1391–1408, 2008a.

M. W. Spratling. Reconciling predictive coding and biased competition models of cortical function. Frontiers in computational neuroscience, 2:4, 2008b.

M. W. Spratling. A review of predictive coding algorithms. Brain and cognition, 112:92–97, 2017.

M. W. Spratling, K. De Meyer, and R. Kompass. Unsupervised learning of overlapping image components using divisive input modulation. Computational intelligence and neuroscience, 2009, 2009.

M. V. Srinivasan, S. B. Laughlin, and A. Dubs. Predictive coding: a fresh view of inhibition in the retina. Proceedings of the Royal Society of London. Series B. Biological Sciences, 216(1205):427–459, 1982.

R. B. Stein, E. R. Gossen, and K. E. Jones. Neuronal variability: noise or part of the signal? Nature Reviews Neuroscience, 6(5):389–397, 2005.

M. Steriade, D. A. McCormick, and T. J. Sejnowski. Thalamocortical oscillations in the sleeping and aroused brain. Science, 262(5134):679–685, 1993.

C. Tallon-Baudry and O. Bertrand. Oscillatory gamma activity in humans and its role in object representation. Trends in cognitive sciences, 3(4):151–162, 1999.

C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Pernier. Stimulus specificity of phase-locked and non-phase-locked 40 hz visual responses in human. Journal of Neuroscience, 16(13):4240–4249, 1996.

A. M. Treisman and G. Gelade. A feature-integration theory of attention. Cognitive psychology, 12(1):97–136, 1980.

I. van de Vijver and M. X. Cohen. Electrophysiological phase synchrony in distributed brain networks as a promising tool in the study of cognition. In New Methods in Cognitive Psychology, pages 214–244. Routledge, 2019.

D. C. Van Essen, C. H. Anderson, and D. J. Felleman. Information processing in the primate visual system: an integrated systems perspective. Science, 255(5043):419–423, 1992.

T. Van Kerkoerle, M. W. Self, B. Dagnino, M.-A. Gariel-Mathis, J. Poort, C. Van Der Togt, and P. R. Roelfsema. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proceedings of the National Academy of Sciences, 111 (40):14332–14341, 2014.

R. VanRullen. Perceptual cycles. Trends in cognitive sciences, 20(10):723–735, 2016.

R. VanRullen and J. Dubois. The psychophysics of brain rhythms. Frontiers in psychology, 2:203, 2011.

R. VanRullen and J. S. Macdonald. Perceptual echoes at 10 hz in the human brain. Current biology, 22(11):995–999, 2012.

R. VanRullen, T. Carlson, and P. Cavanagh. The blinking spotlight of attention. Proceedings of the National Academy of Sciences, 104(49):19204–19209, 2007.

S. V. Vaseghi. Advanced Digital Signal Processing and Noise Reduction. John Wiley and Sons Ltd, 2000.

M. Vinck, M. van Wingerden, T. Womelsdorf, P. Fries, and C. M. Pennartz. The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization. Neuroimage, 51(1):112–122, 2010.

P. Wallisch and J. A. Movshon. Structure and function come unglued in the visual cortex. Neuron, 60(2):195–197, 2008. ISSN 0896-6273.

E. Watanabe, A. Kitaoka, K. Sakamoto, M. Yasugi, and K. Tanaka. Illusory motion reproduced by deep neural networks trained for prediction. Frontiers in psychology, 9:345, 2018.

S. Watanabe. Information-theoretical aspects of inductive and deductive inference. IBM journal of research and development, 4(2):208–231, 1960.

V. Weilnhammer, H. Stuke, G. Hesselmann, P. Sterzer, and K. Schmack. A predictive coding account of bistable perception-a model-based fmri study. PLoS computational biology, 13 (5):e1005536, 2017.

H. Wen, K. Han, J. Shi, Y. Zhang, E. Culurciello, and Z. Liu. Deep predictive coding network for object recognition. In International Conference on Machine Learning, pages 5266–5275. PMLR, 2018.

J. M. Wolfe, M. L.-H. Võ, K. K. Evans, and M. R. Greene. Visual search in scenes involves selective and nonselective pathways. Trends in cognitive sciences, 15(2):77–84, 2011.

G. F. Woodman. A brief introduction to the use of event-related potentials in studies of perception and attention. Attention, Perception, & Psychophysics, 72(8):2031–2046, 2010.

V. Wyart and C. Tallon-Baudry. Neural dissociation between visual awareness and spatial attention. Journal of Neuroscience, 28(10):2667–2679, 2008.

D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3):356–365, 2016.

H. Zhang, A. J. Watrous, A. Patel, and J. Jacobs. Theta and alpha oscillations are traveling waves in the human neocortex. Neuron, 98(6):1269–1281, 2018.