THÉORIES SCIENTIFIQUES DE LA CONSCIENCE

Séance 7

Jonathan Simon

PHI-6385

APERÇU

- 1) Le problème de la méthodologie pour les cas faciles
- 2) Le problème de la méthodologie pour les cas difficiles
- 3) Aperçu des théories et de ce qu'elles impliquent au sujet de la conscience dans l'IA

LE PROBLÈME DE LA MÉTHODOLOGIE POUR LES CAS FACILES

MÉTHODOLOGIE POUR LES CAS FACILES

- Comment trouver les corrélats neuronaux (ou fonctionnels) de la conscience - en toi-même ?
- 1) Prends des notes, rapporte et documente ce que tu vis et à quel moment.
- 2) Fais des scanners cérébraux sur toi-même en même temps.
- 3) compare-le

MÉTHODOLOGIE POUR LES CAS FACILES

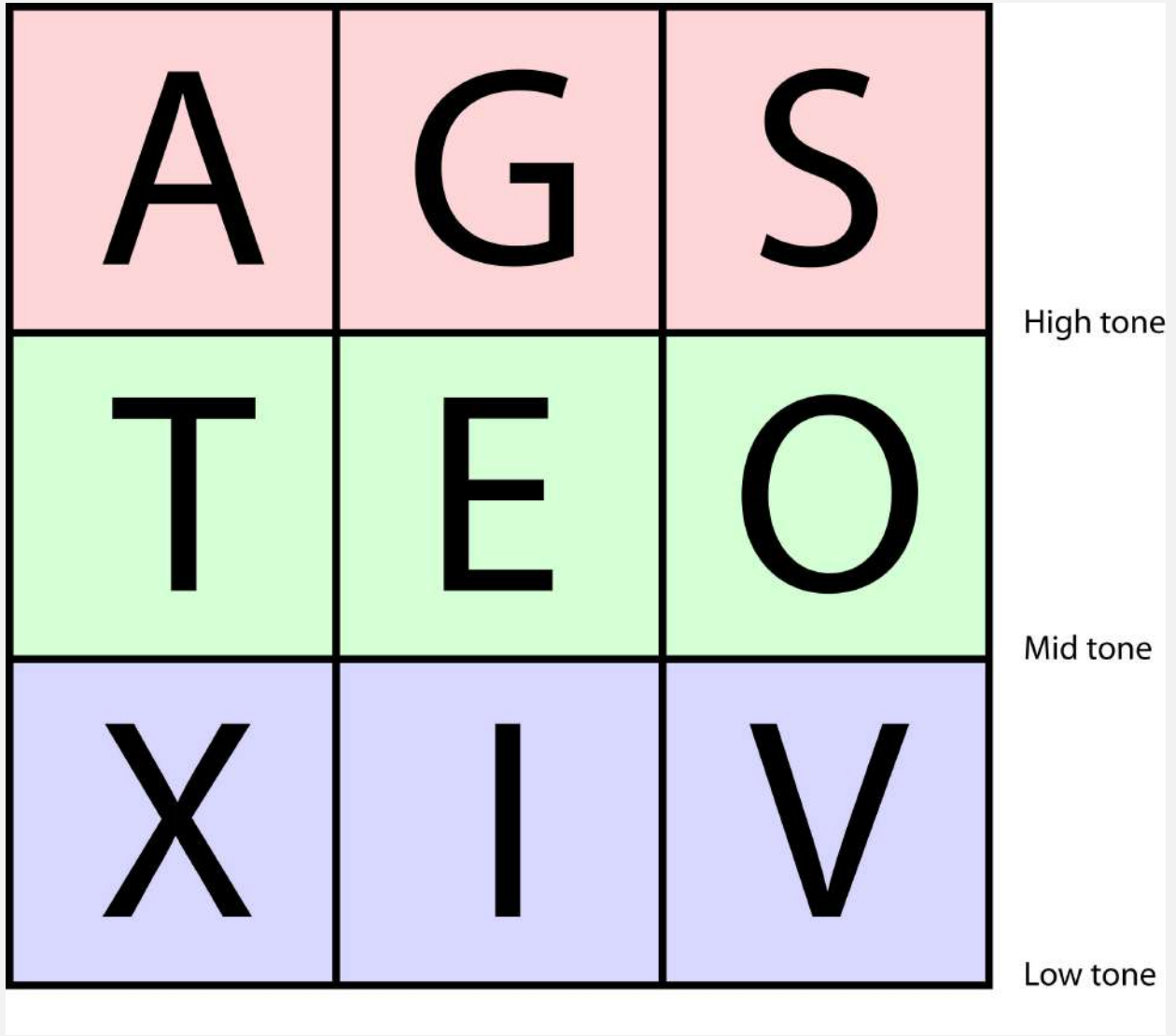
- Génial ! Mais :
- a) comment étendre tes découvertes à d'autres personnes ? (le problème des autres esprits)
- b) comment être certain que tes rapports sont corrects (le problème de la mémoire) ?
- c) l'illusion de la lumière du réfrigérateur : cette méthode ne fonctionne que pour les expériences dont tu peux rendre compte / faire un rapport. Mais i) que se passe-t-il si ton rapport sur l'état le biaise ? ii) que se passe-t-il s'il y a des états dont tu peux faire l'expérience mais que tu ne peux pas rapporter ?

MÉTHODOLOGIE POUR LES CAS FACILES

- c) l'illusion de la lumière du réfrigérateur

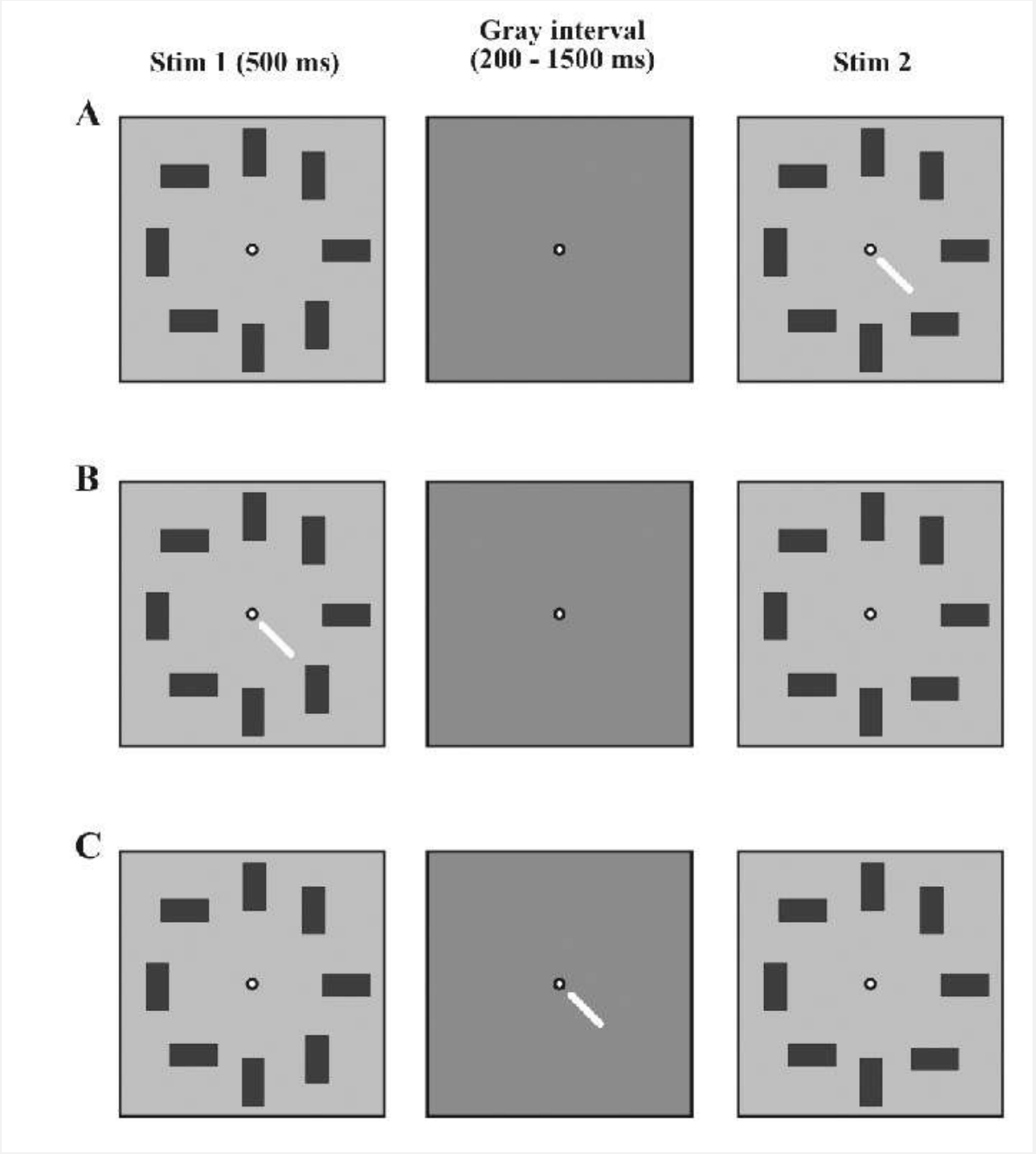
- Comment évaluer si nos mécanismes de signalement sont des parties constitutives des CCN de la conscience, ou s'ils sont indépendants, mais nécessaires pour que nous puissions parler de notre conscience ?

DÉBORDEMENT : SPERLING 1960



DÉBORDEMENT : CÉCITÉ AU CHANGEMENT?

DÉBORDEMENT : LANDMAN 2003



MÉTHODOLOGIE POUR LES CAS FACILES

- En raison de ces complications, la science de la conscience devient très théorique :
- le paradigm «no-report»
- des marqueurs neuronaux spécifiques de la conscience comme l'onde P300
- la recherche de types neuronaux ou psychologiques «naturels» (natural kinds).

MÉTHODOLOGIE POUR LES CAS FACILES

- Invariablement, il y a un désaccord (philosophique) sur la façon de trouver des marqueurs indépendants des rapports, et plus généralement sur la façon de dériver des prédictions à partir de théories

The New York Times

ORIGINS

2 Leading Theories of Consciousness Square Off

Scientists revealed the results of experiments testing how our brains give rise to conscious thought — and ended a 25-year-old bet.

ARC-COGITATE

- Une compétition adversarielle entre deux théories de la conscience (Espace de travail global et Théorie de l'information intégrée), dans laquelle les deux parties se sont mises d'accord à l'avance sur les expériences - et sur les résultats qui soutiendraient les deux parties dans ces expériences.

NATURE BRIEFING | 21 September 2023

Daily briefing: Critics call consciousness theory ‘pseudoscience’

A group of researchers say that a high-profile theory about consciousness is receiving undue attention and can’t be empirically tested. Plus, world leaders have pledged to redouble their efforts towards the Sustainable Development Goals and ancient whittled logs could be the earliest known wooden structure.

Flora Graham

LE PROBLÈME DE LA MÉTHODOLOGIE POUR LES CAS DIFFICILES

MÉTHODOLOGIE POUR LES CAS DIFFICILES

- (De la dernière séance) :
- 1) Les mesures comportementales sont faciles à tromper (surtout pour les IA).
- 2) Les "marqueurs" du cas humain / bio, même si nous pouvions nous mettre d'accord sur eux, pourraient n'être que des accidents de mise en œuvre (comment nous exécutons une certaine fonction).
- 3) Il est difficile de décider quelle théorie est correcte (voir la section précédente).

MÉTHODOLOGIE POUR LES CAS DIFFICILES

- 4) Le problème des petits reseaux

LE PROBLÈME DES PETITS RESEAUX

- Exemples :

- **Théorie de l'espace de travail global :**

Un nœud centralisé qui transmet à une majorité d'autres nœuds.

LE PROBLÈME DES PETITS RESEAUX

Exemples:

- **Théorie de la pensée d'ordre supérieur :**

  Un nœud qui suit/réfléchit le contenu d'un autre nœud.

LE PROBLÈME DES PETITS RESEAUX

- Exemples :

- **Théorie de la mémoire fragile à court terme :**

Contenu représenté sur plusieurs couches d'un réseau récurrent.

LE PROBLÈME DES PETITS RESEAUX

Exemples:

Théorie de la réafférence / du tronc cérébral :

Un nœud qui permet de savoir si les changements apportés à un autre nœud sont endogènes ou exogènes.

NOTRE APPROCHE

- Bayesian Theory-Heavy Approach

- 1) Tu n'as pas besoin de choisir une théorie, il te suffit de prendre ta distribution de crédibilité (quelle est la probabilité que chaque théorie soit vraie).

- Nous dérivons ensuite des indicateurs, où la probabilité qu'une chose soit consciente si elle possède tous ces indicateurs, étant donné que la théorie en question est vraie, devrait être relativement élevée

NOTRE APPROCHE

- Bayesian Theory-Heavy Approach

- 2) Bien que ce ne soit pas l'objet de notre rapport, nous supposons également que plus un système satisfait d'indicateurs, moins le problème du petit réseau est pressant

APERÇU DES THÉORIES ET DE CE QU'ELLES IMPLIQUENT AU SUJET DE LA CONSCIENCE DANS L'IA

Recurrent processing theory
RPT-1: Input modules using algorithmic recurrence
RPT-2: Input modules generating organised, integrated perceptual representations
Global workspace theory
GWT-1: Multiple specialised systems capable of operating in parallel (modules)
GWT-2: Limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism
GWT-3: Global broadcast: availability of information in the workspace to all modules
GWT-4: State-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks
Computational higher-order theories
HOT-1: Generative, top-down or noisy perception modules
HOT-2: Metacognitive monitoring distinguishing reliable perceptual representations from noise
HOT-3: Agency guided by a general belief-formation and action selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring
HOT-4: Sparse and smooth coding generating a "quality space"
Attention schema theory
AST-1: A predictive model representing and enabling control over the current state of attention
Predictive processing
PP-1: Input modules using predictive coding
Agency and embodiment
AE-1: Agency: Learning from feedback and selecting outputs so as to pursue goals, especially where this involves flexible responsiveness to competing goals
AE-2: Embodiment: Modeling output-input contingencies, including some systematic effects, and using this model in perception or control