

Introspection et métacognition :
Les mécanismes de la connaissance de soi

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

Cours

Liens entre métacognition et théorie de l'esprit


Liens entre métacognition et « théorie de l'esprit »

- La métacognition implique de se représenter son propre esprit en train de représenter une information
 - Par exemple:
 - une voiture rouge est passée ce matin
 - Conscience primaire: j'ai vu une voiture rouge
 - Mémoire: je me souviens que (j'ai vu une voiture rouge)
 - Méta-mémoire: je sais que (je me souviens que (j'ai vu une voiture rouge))
 - Un exemple plus intéressant étudié la semaine dernière:
 - Grace à la psychologie expérimentale , nous découvrons les mécanismes non-conscients de détection d'erreurs.
 - Conclusion: mon cerveau détecte inconsciemment mes erreurs.
 - je ne sais pas que (je sais que (j'ai fait une erreur))
- Ces méta-représentations doivent avoir un format de codage très similaire, qu'elles portent sur notre propre esprit ou sur celui de quelqu'un autre:
 - Je sais que **tu** ne sais pas qqc versus Je sais que **je** ne sais pas qqc
 - Dans les deux cas, il faut spécifier l'agent (moi ou un autre), l'attitude mentale (croire, savoir...), et la proposition examinée.
- Il se pourrait donc que nous utilisions le même format de représentation mentale et les mêmes aires cérébrales pour représenter notre esprit et celui des autres.

Arguments en faveur d'un lien étroit entre métacognition et théorie de l'esprit

- La connaissance de soi et la connaissance de l'autre se développent simultanément chez l'enfant
- Ces deux types de connaissances ne sont pas indépendantes, mais interagissent entre elles:
 - interférence au cours d'un essai (Kovacs et al., *Science* 2010)
 - généralisation au cours de l'apprentissage (Melztoff et Brooks, *Developmental Psychology* 2008)
- Elles font appel à un réseau similaire d'aires cérébrales
- En particulier, il existerait une représentation cérébrale partagée pour nos propres erreurs et celles d'une tierce personne.

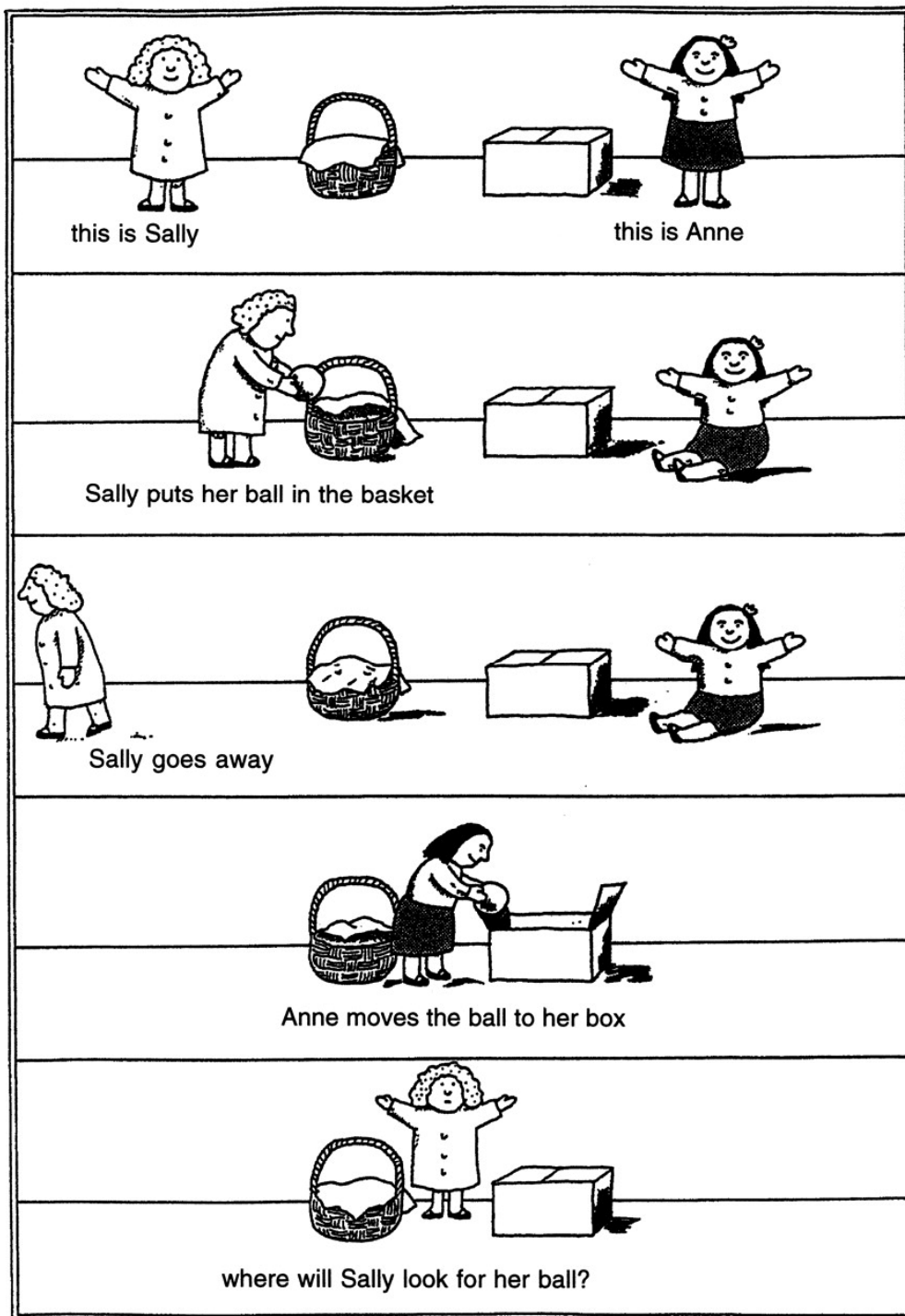
La « théorie de l'esprit »

- Capacité de se représenter les autres personnes comme:
 - des êtres pensants
 - dotés d'intentions et de croyances
 - dont les pensées peuvent différer
 - (1) de la réalité
 - (2) de mes propres croyances.
- Le concept d'attitude ou “posture intentionnelle” (*intentional stance*), introduit par Franz Brentano puis Daniel Dennett
 - Travaux pionniers de Fritz Heider sur la perception sociale et la théorie de l'attribution 
- Le terme de “théorie de l'esprit”, introduit par David Premack

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences* 4, 515-526.
- Méthodologie essentielle: la tâche de la croyance erronée ou tâche de *Sally et Anne*

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103-128.

voir la méta-analyse de Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev*, 72(3), 655-684.



- Pour réussir ce test, l'enfant doit dissocier ses propres croyances de celle des autres protagonistes
- Les enfants réussissent à 4 ans, mais pas à 3 ans
- Ces résultats sont répliqués par Uta Frith qui, avec Alan Leslie et Simon Baron-Cohen, introduit l'hypothèse que l'**autisme** correspond à un déficit sélectif de la théorie de l'esprit.

Leslie, A.M. & Frith, U. (1988). Autistic children's understanding of seeing, knowing and believing. *British Journal of Developmental Psychology* **6**: 315–324.

Baron-Cohen S, Leslie AM, Frith U (1985). Does the autistic child have a 'theory of mind'? *Cognition* **21** (1): 37–46.

C D Frith, U Frith Science 1999;286:1692-1695



Liens avec la méta-cognition?

- C'est au même âge que les enfants commencent à
 - faire la différence entre l'apparence et la réalité
 - montrer une représentation métacognitive de leur propre compétence
 - comprendre l'esprit des autres
- Exemple: le test des smarties

voir par exemple Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Dev*, 59(1), 26-37.

L'enfant voit une boîte de smarties. Il croit qu'elle est pleine de smarties.

On lui montre qu'elle contient en fait un crayon.

Question: qu'y a-t-il dans la boîte? Réponse: un crayon

Qu'est-ce que tu croyais qu'il y avait dans la boîte?

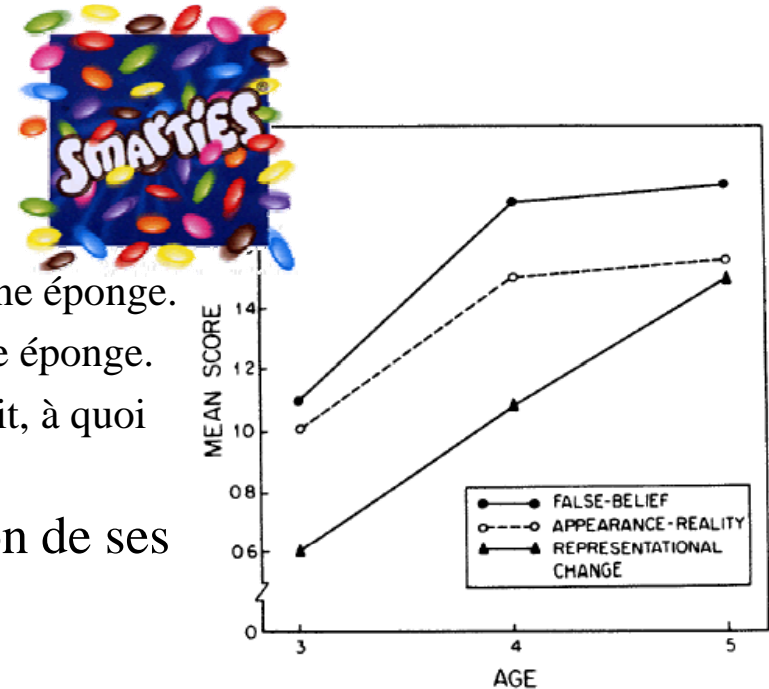
Réponse: un crayon

Dans un autre test, ce qui ressemble à un rocher s'avère être une éponge.

Les enfants de trois ans disent avoir toujours su que c'était une éponge.

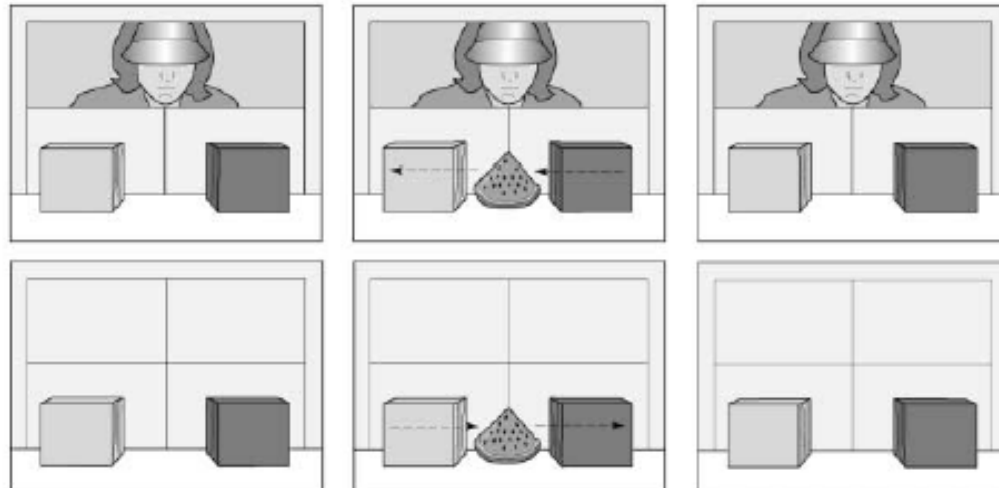
D'autres questions portent sur ce que quelqu'un d'autre croirait, à quoi ressemble l'objet, et ce qu'il est réellement.

Les trois tests sont corrélés. La méta-représentation de ses propres connaissances se développe en même temps que celle des connaissances des autres.



Développement précoce de la théorie de l'esprit

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255-258.



Test trial

Green-box condition



Yellow-box condition

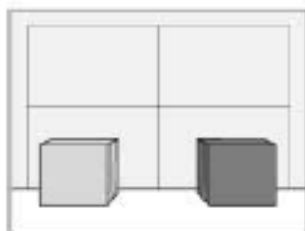


Fig. 3. Events shown during the test trial.

Onishi et Baillargeon développent une nouvelle version non-verbale du test:

- L'enfant voit une vidéo qui représente l'acquisition d'une croyance vraie ou fausse de la part de l'expérimentateur.
- Puis il voit la personne rechercher l'objet dans l'une ou l'autre des cachettes.

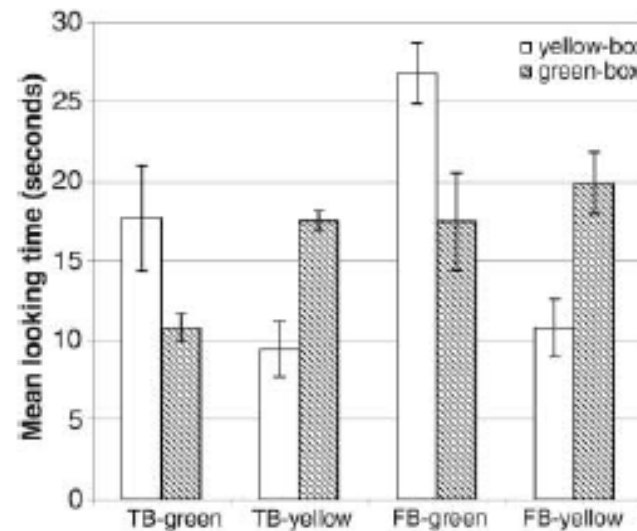


Fig. 4. Mean (\pm SE) looking times during the test trial (after the actor reached into the green or yellow box) in the four belief conditions.

L'enfant de 15 mois est surpris et regarde plus longtemps lorsque la personne cherche à l'endroit inapproprié selon sa croyance à elle (et non selon la connaissance de l'enfant).

Cela implique une connaissance implicite de ce que l'autre personne sait.

Développement précoce de la théorie de l'esprit

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255-258.

A 15 mois, une capacité de se représenter la pensée d'autrui existerait donc déjà.

Une interprétation alternative est présentée par J. Perner, T. Ruffman, Infants' insight into the mind: How deep? *Science*, 308, 214-216 (2005).

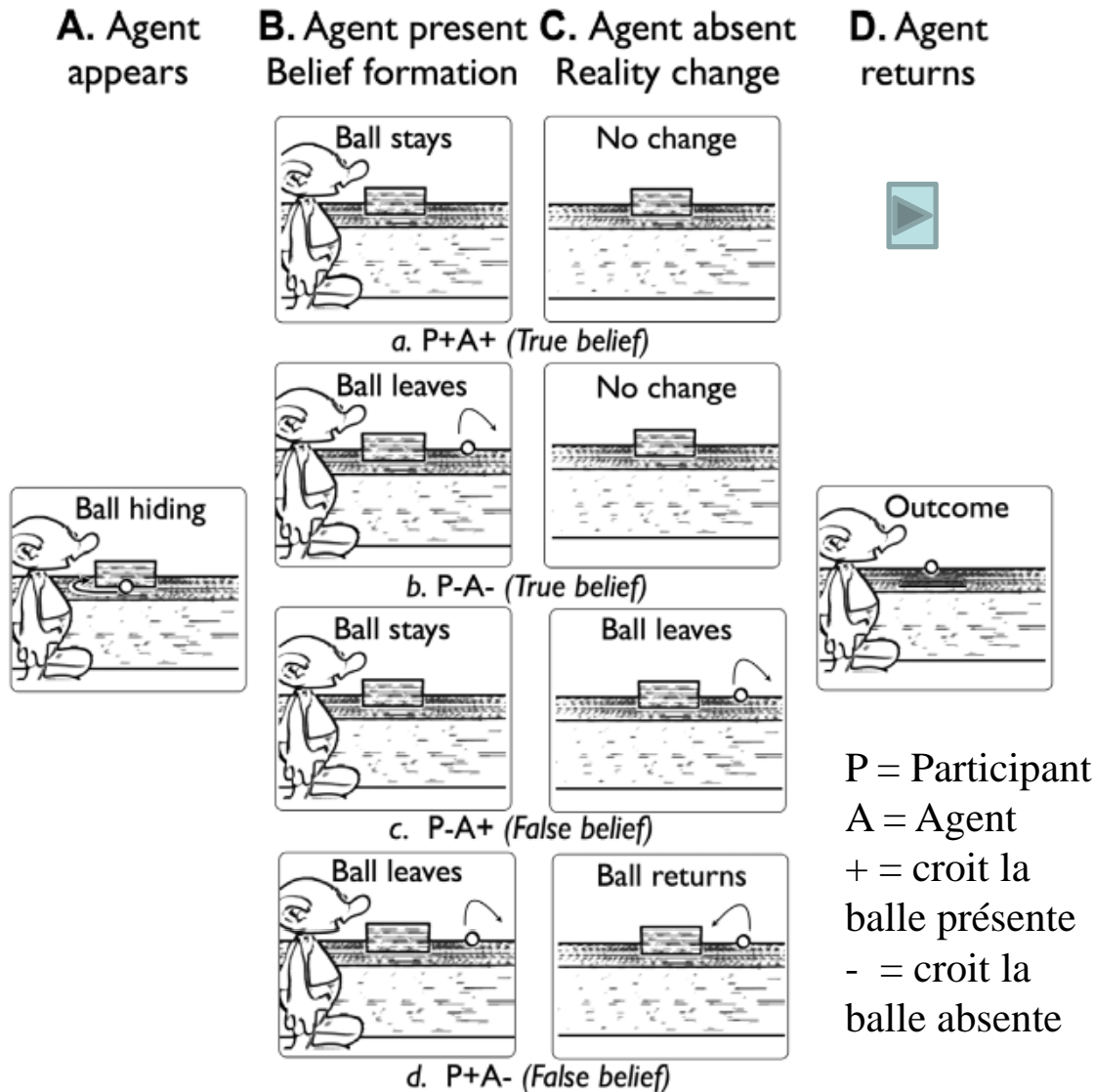
Mais les résultats sont répliqués et étendus par plusieurs chercheurs:

- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337-342.
- Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends Cogn Sci*, 14(3), 110-118.
- Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Dev*, 80(4), 1172-1196.
- Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Dev Sci*, 13(6), 907-912.
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychol Sci*, 18(7), 587-592.
- Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychol Sci*, 18(7), 580-586.
- Kovacs, A. M., Teglas, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, 330, 1830-1834.

Interférence entre connaissances personnelles et connaissances d'autrui chez l'adulte et le bébé de 7 mois

Kovacs, A. M., Teglas, E., & Endress, A. D. (2010).

The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, 330, 1830-1834.



Questions posées:

- La représentation des connaissances d'autrui est-elle automatique, même lorsqu'elle n'est pas nécessaire à la tâche?
- Est-elle dans le même format perceptif que la connaissance personnelle?

Tâche demandée aux adultes:

- Examiner un film décrivant une croyance vraie ou fausse, sur la position d'une balle

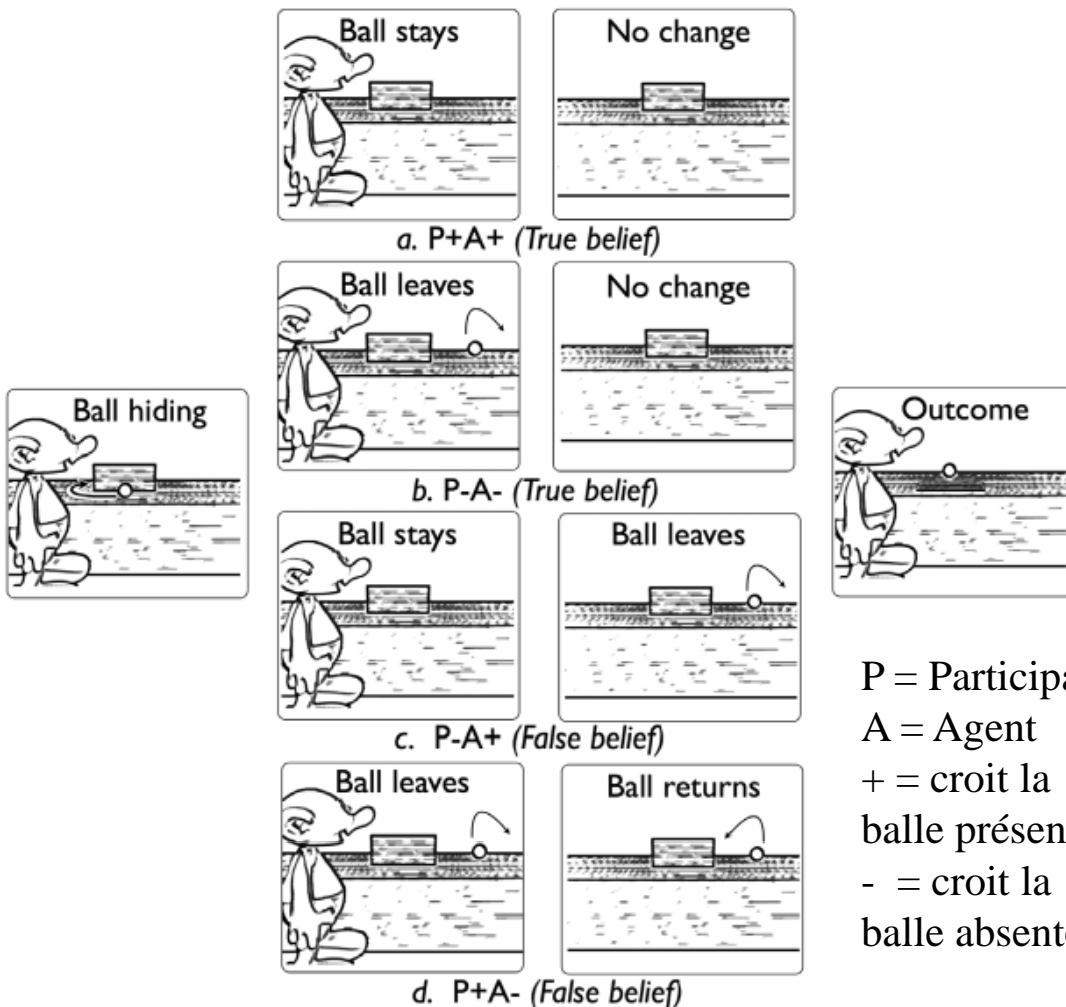
- Détecter le plus vite possible si une balle est présente ou non lorsque l'écran s'abaisse (quel que soit le film précédent).

Interférence entre connaissances personnelles et connaissances d'autrui chez l'adulte et le bébé de 7 mois

Kovacs, A. M., Teglas, E., & Endress, A. D. (2010).

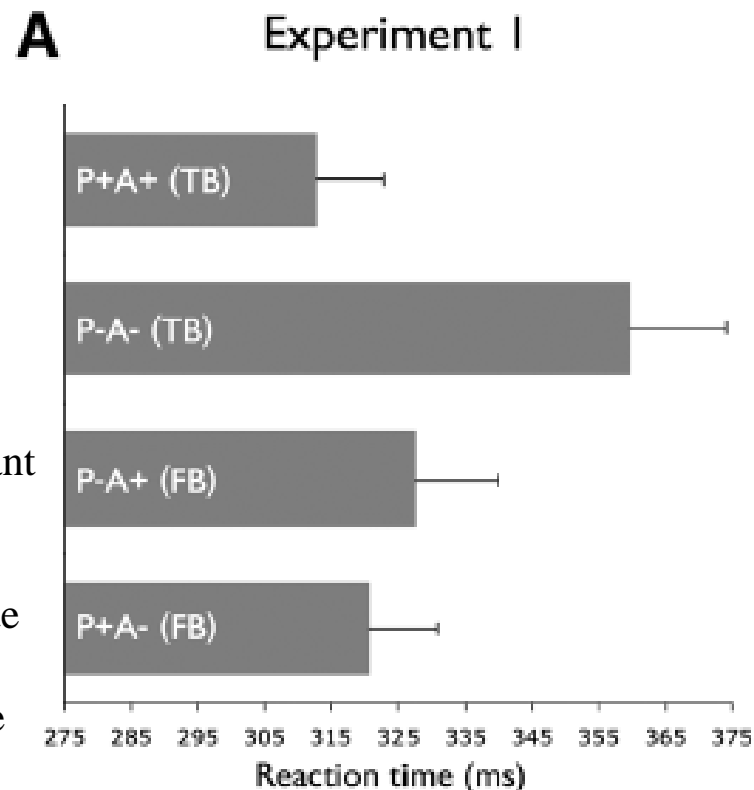
The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, 330, 1830-1834.

A. Agent appears **B. Agent present** **C. Agent absent** **D. Agent returns**
 Belief formation Reality change



Résultats:

- accélération du temps de réaction lorsque le participant sait qu'une balle est cachée ou, crucialement, lorsqu'il sait qu'un autre croit qu'une balle est cachée.

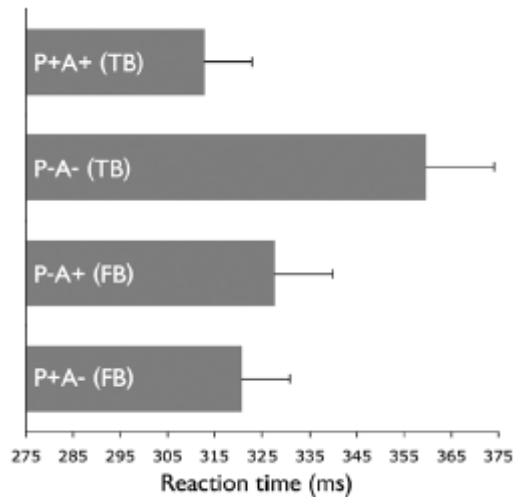


Interférence entre connaissances personnelles et connaissances d'autrui chez l'adulte et le bébé de 7 mois

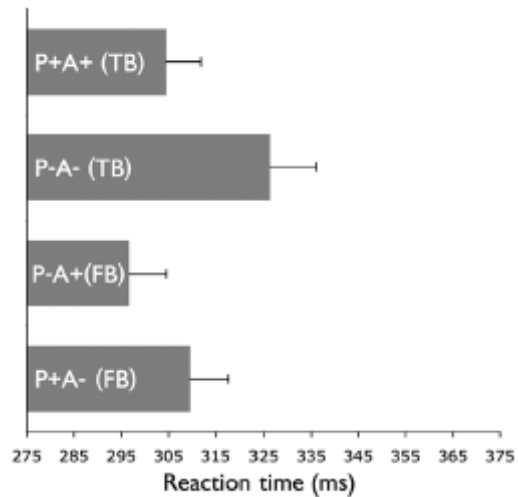
Kovacs, A. M., Teglas, E., & Endress, A. D. (2010).

The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, 330, 1830-1834.

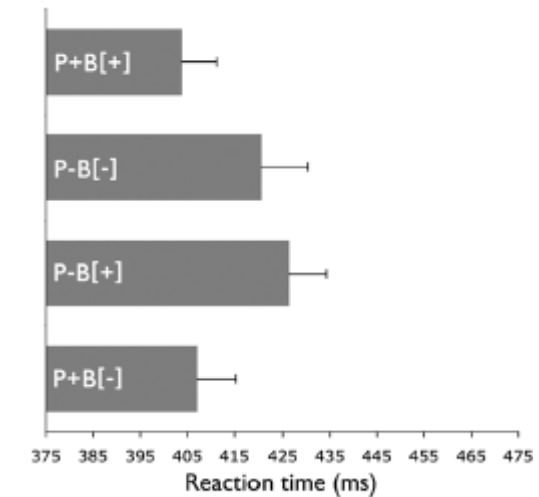
Exp 1: l'observateur revient à la fin du film.



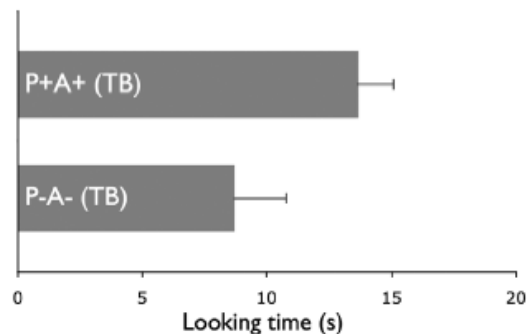
Exp 2: l'observateur ne revient pas à la fin du film.



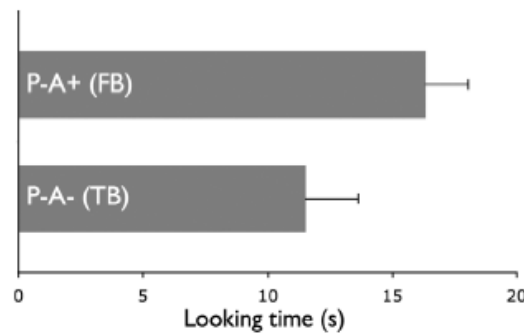
Exp 3: l'observateur est remplacé par une pile d'objets



A Experiment 4. True belief



B Experiment 5. False belief



Les expériences sont ensuite étendues à l'enfant de 7 mois, avec pour mesure la surprise (durée du regard).

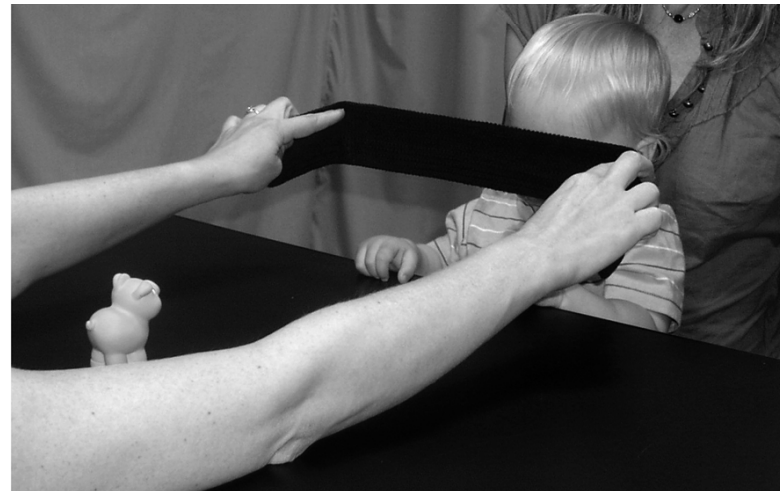
→ dès cet âge, les enfants sont surpris de l'absence d'une balle attendue ou, crucialement, dont l'enfant croit que l'autre personne l'attend.

→ dès cet âge, existerait un système de représentation partagé pour soi et autrui.

Généralisation de la connaissance de soi à la connaissance de l'autre chez l'enfant de 12 mois

(Meltzoff & Brooks, Developmental Psychology 2008)

- Quelles sont les métaconnaissances des enfants sur la vision?
- Les enfants de 10 et 12 mois suivent un adulte lorsqu'il tourne la tête avec les yeux ouverts, mais pas avec les yeux fermés (Brooks & Meltzoff, 2002).
- Par contre, ils suivent également une personne qui porte un bandeau
- Explication? Peut-être les enfants se reposent-ils sur leur propre expérience de fermer les yeux (alors qu'ils n'ont pas d'expérience personnelle avec un bandeau)
- Nouvelle expérience: Des enfants de 12 mois sont incités à jouer avec des objets alors qu'un bandeau est parfois interposé devant eux. Dans le groupe contrôle, le bandeau possède une ouverture et n'empêche pas de voir.



Généralisation de la connaissance de soi à la connaissance de l'autre chez l'enfant de 12 mois

(Meltzoff & Brooks, Developmental Psychology 2008)

- Résultats:

Après l'entraînement, seuls les enfants entraînés avec le bandeau opaque cessent de suivre l'orientation de la tête d'un adulte qui a les yeux bandés.

- Expérience 2:

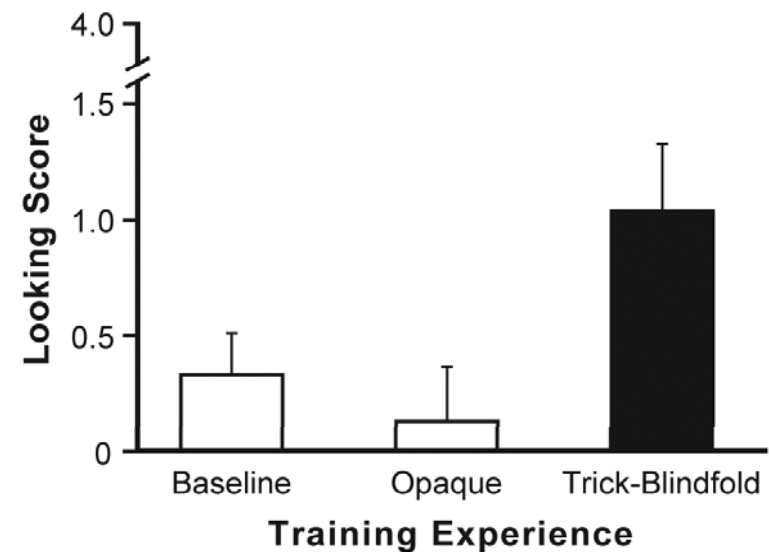
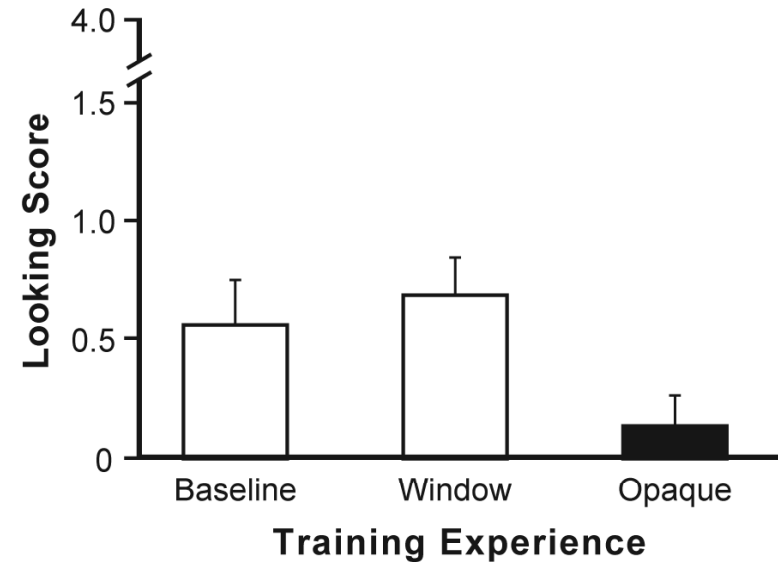
A l'inverse, des enfants de 18 mois ne suivent plus un adulte qui a les yeux bandés (Brooks & Meltzoff, 2002).

Ce comportement peut-il également être modifié?

Entraînement des enfants avec un bandeau transparent!

Cet entraînement restaure un comportement de suivi de l'orientation de la tête d'un adulte qui a les yeux bandés.

Conclusion: la métacognition de soi et la théorie de l'esprit des autres partagent des représentations communes.



Les bases cérébrales de la théorie de l'esprit

Un réseau comprenant le cortex préfrontal antéro-mésial, le précuneus, la jonction temporo-pariétale (particulièrement à droite) et la partie antérieure du lobe temporal est impliqué dans la théorie de l'esprit et notamment les tâches de fausse croyance.

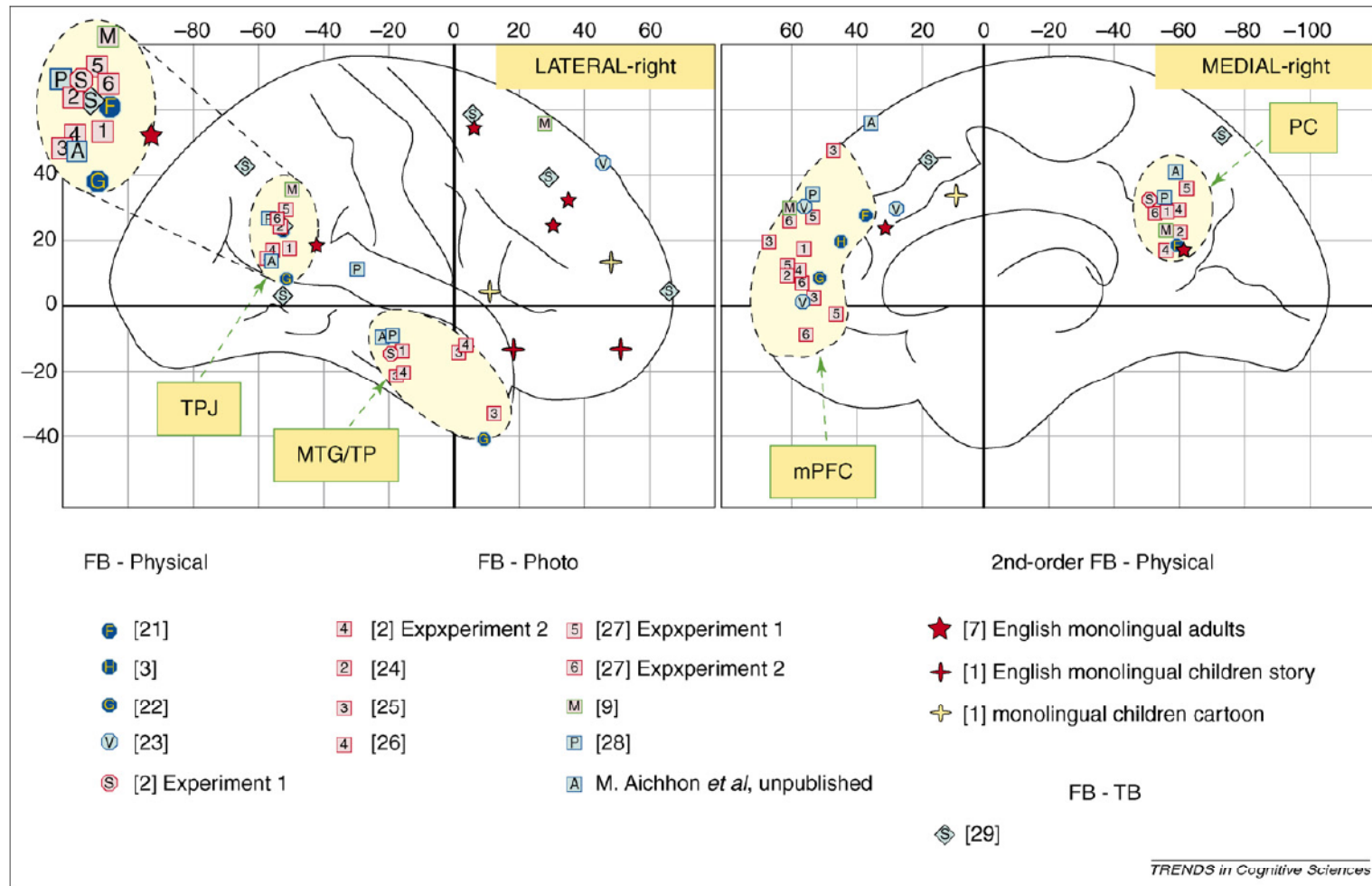


Figure 1. Location of peak voxels in y- and z-space coordinates (outlines of cerebral structures are only approximate) in three lateral and medial areas reported in imaging studies of false-belief (FB) processing in comparison with data reported by Kobayashi *et al.* [1,7] for monolingual English speakers. Abbreviations: mPFC, medial prefrontal cortex; MTG, middle temporal gyrus; PC, precuneus; TP, temporal pole; TPJ, temporo-parietal junction.

Théorie de l'esprit: la représentation de la pensée d'autrui

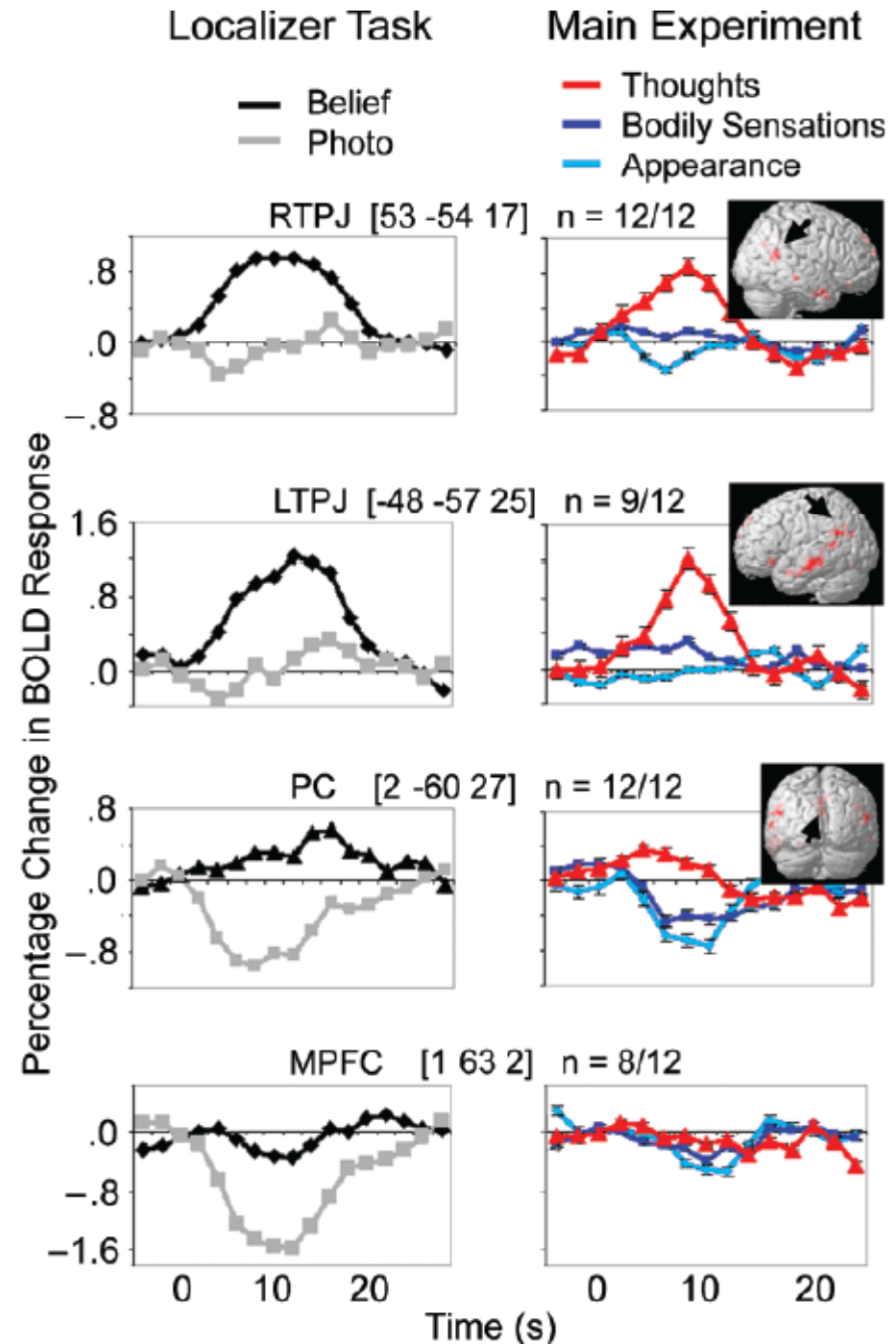
Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychol Sci*, 17(8), 692-699.

Les travaux de Rebecca Saxe montrent que l'on peut activer ce réseau en écoutant une histoire qui porte sur les états mentaux d'autrui, par opposition à leurs sensations ou leur aspect physique.

“Joe was a heavy-set man, with a gut that fell over his belt. He was balding and combed his blonde hair over the top of his head. His face was pleasant, with large brown eyes.”

versus

“Nicky knew that his sister’s flight from San Francisco was delayed ten hours. Only one flight was delayed so much that night, so when he got to the airport, he knew that flight was hers.”



La pensée d'autrui versus mes propres pensées

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., et al. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage*, 14(1 Pt 1), 170-181.

Dessin expérimental 2x2:

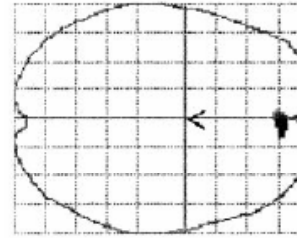
Histoires parlant soit

- de faits décousus,
- de l'état d'esprit d'autrui
- du propre état d'esprit de la personne scannée
- ou des deux

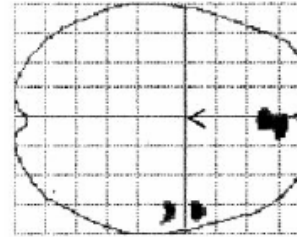
Exemple

« You went to London for a weekend trip and you would like to visit some museums and different parks around London. In the morning, when you leave the hotel, the sky is blue and the sun is shining. So you do not expect it to start raining. However, walking around in a big park later, the sky becomes gray and it starts to rain heavily. You forgot your umbrella.

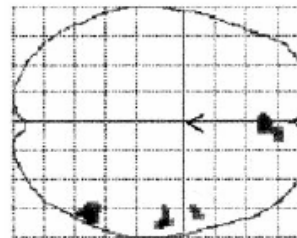
Question: What do you think?»



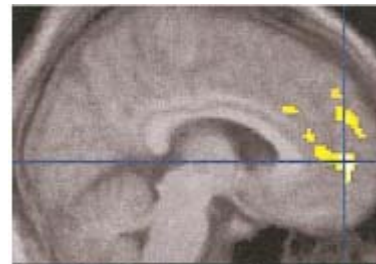
Condition TOM+, SELF-



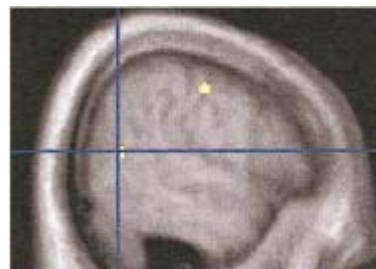
Condition TOM+, SELF+



Condition TOM-, SELF+



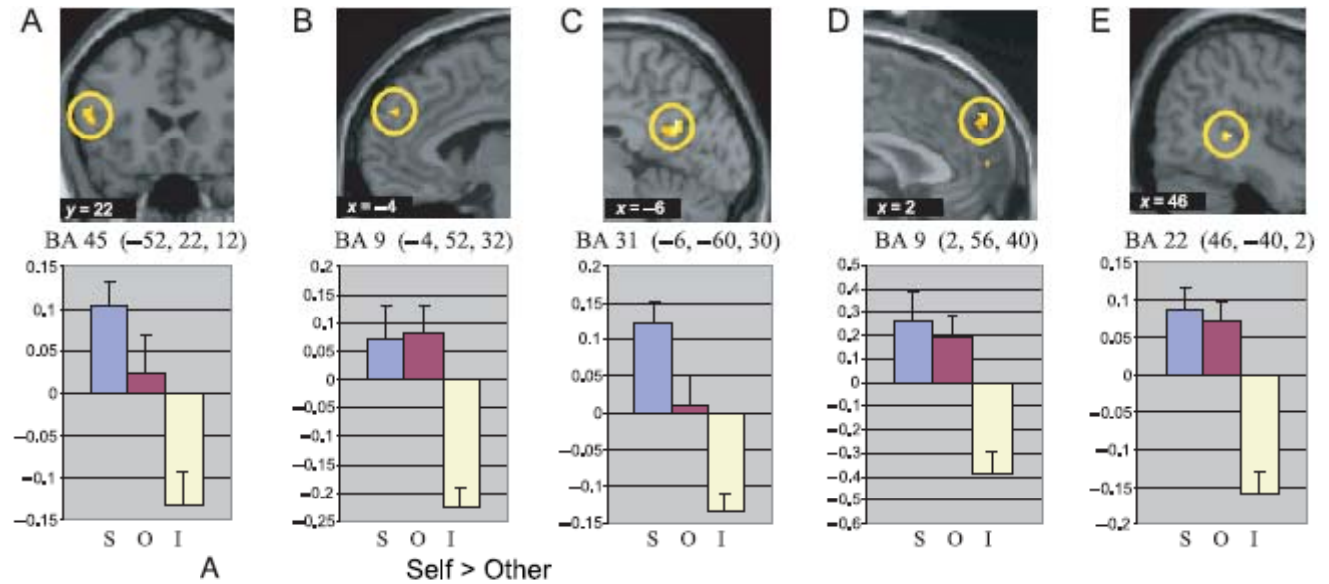
Effet principal de TOM



Effet principal de SELF

La pensée d'autrui versus mes propres pensées

Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., et al. (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *J Cogn Neurosci*, 16(10), 1746-1772.

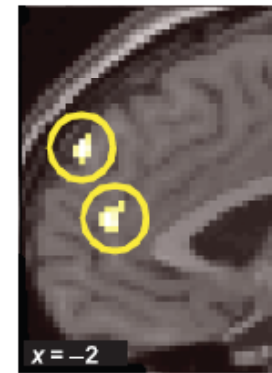
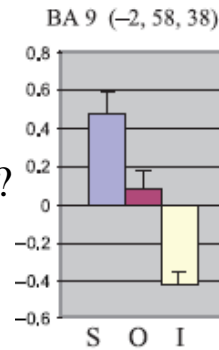


Dessin expérimental:

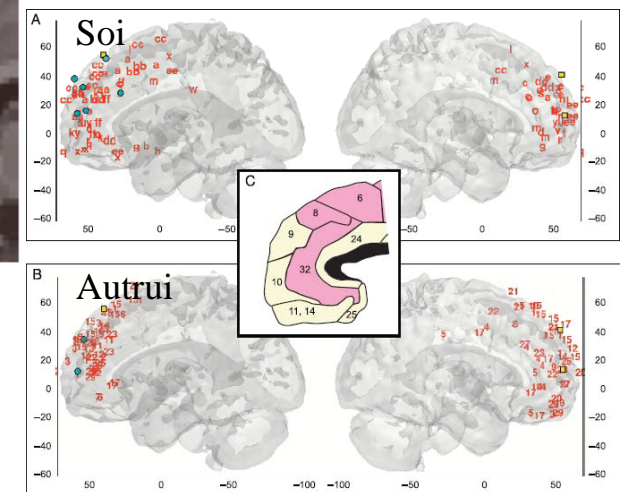
Visualiser des photographies

Trois jugements dans des blocs distincts:

- émotion éprouvée par la personne scannée?
- émotion éprouvée par la personne présente sur la photo?
- contrôle: scène d'extérieur ou d'intérieur?

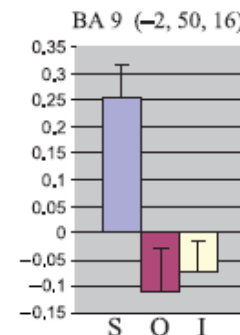


méta-analyse de nombreux travaux:



Résultats:

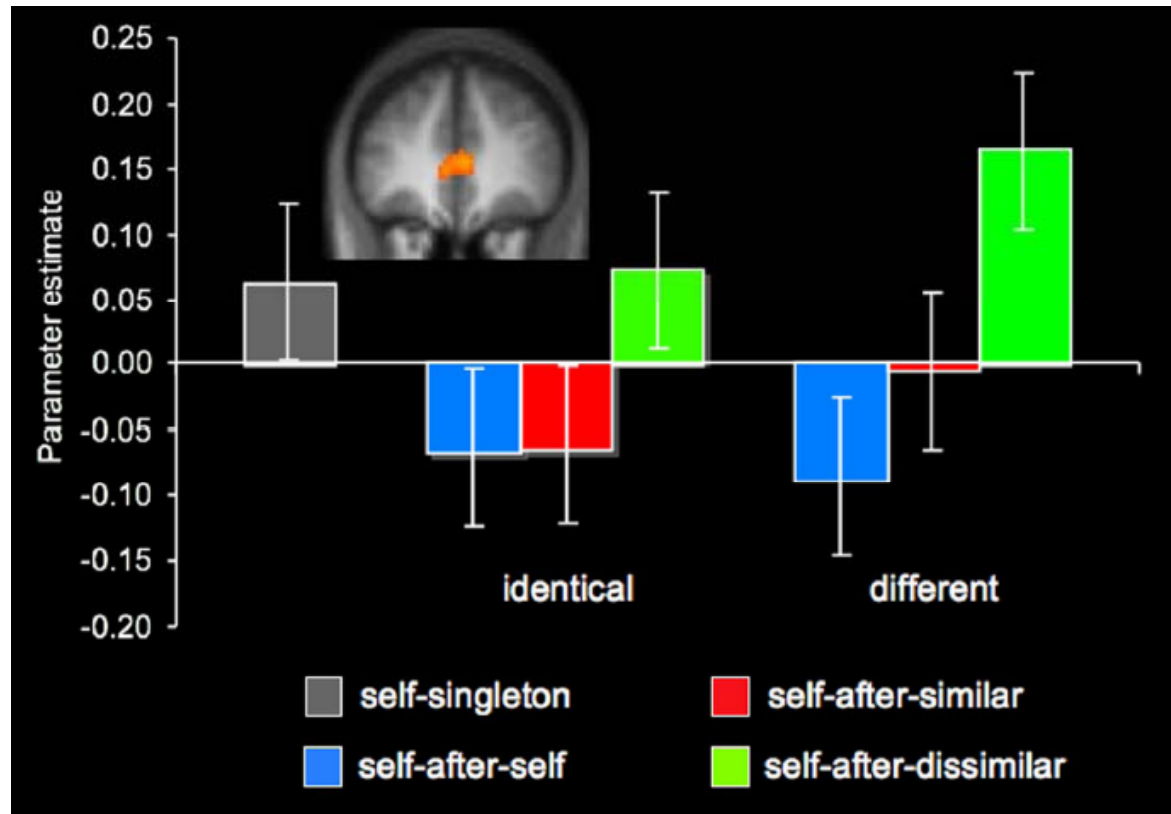
- existence d'un vaste réseau commun (préfrontal mésial, précuneus, temporal postérieur droit)
- mais aussi de régions distinctes du cortex préfrontal mésial, activées pour le soi plus que pour autrui (et vice-versa dans d'autres régions).



La pensée d'autrui versus mes propres pensées

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proc Natl Acad Sci U S A*, 105(11), 4507-4512.

Paradigme de « répétition suppression »: La représentation d'un état mental identique chez soi-même et chez quelqu'un d'autre active exactement la même assemblée de neurones.



Réfléchir à des questions (“enjoy crossword puzzles?”; “like to be the center of attention?”)

Deux questions sont posées à un bref intervalle. La seconde porte sur soi-même; celle qui précède peut

- être identique ou non

- être posée à propos de soi, de quelqu'un d'autre qui partage les mêmes opinions (« similar »), ou qui ne les partage pas (« dissimilar »).

Résultat: dans une région prélocalisée du cortex préfrontal mésial, choisie pour son implication dans la réflexion sur soi-même, il existe une réduction du signal lorsque l'on réfléchit deux fois de suite à soi-même (mais pas nécessairement à la même question), et également lorsque l'on réfléchit à une personne similaire puis à soi-même.

Conclusion proposée: partage des représentations de soi et d'autrui.

Autisme, théorie de l'esprit et représentation de soi

BRAIN
A JOURNAL OF NEUROLOGY

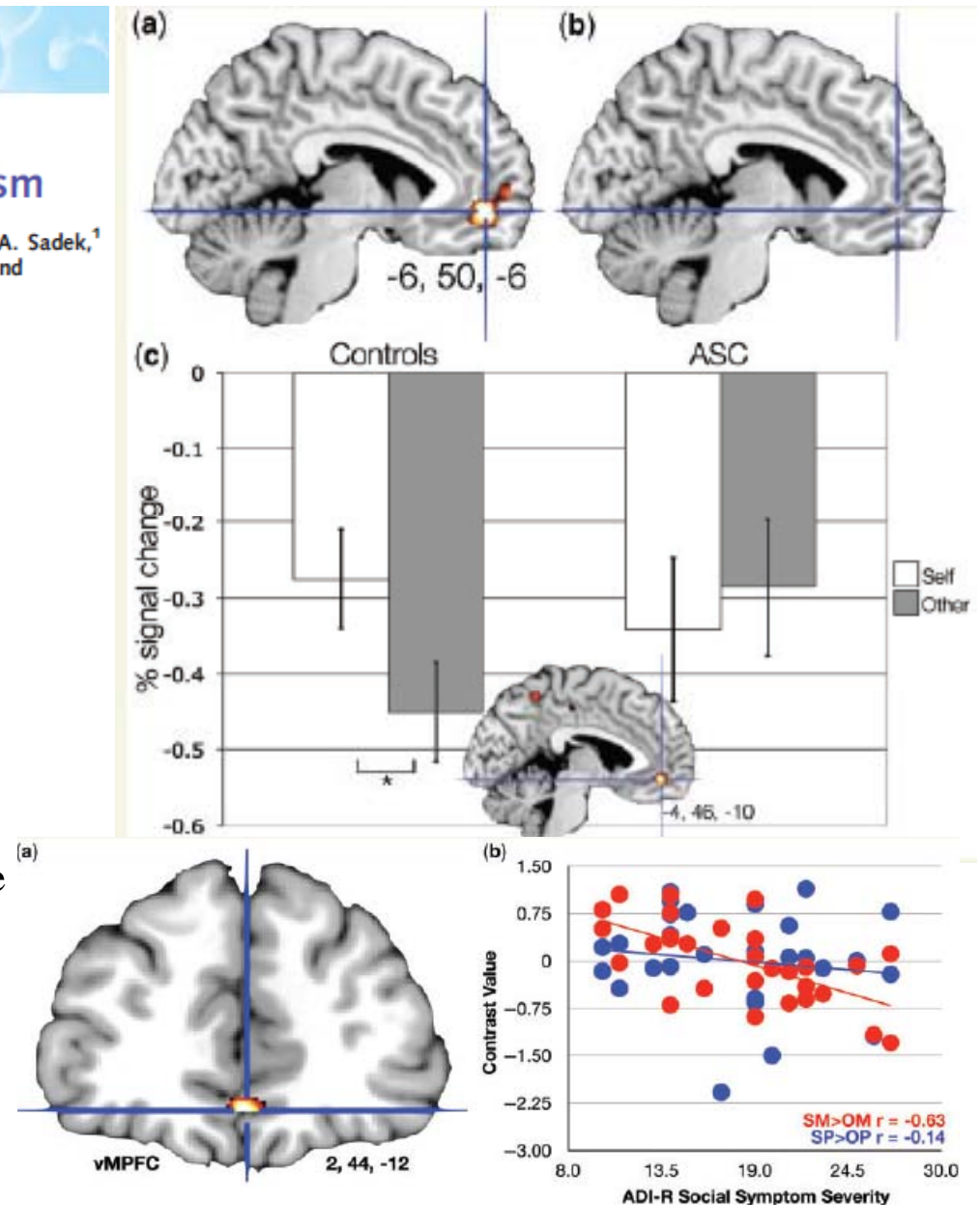
Atypical neural self-representation in autism

Michael V. Lombardo,¹ Bhisudev Chakrabarti,^{1,2} Edward T. Bullmore,³ Susan A. Sadek,¹ Greg Pasco,¹ Sally J. Wheelwright,¹ John Suckling,³ MRC AIMS Consortium* and Simon Baron-Cohen¹

IRM fonctionnelle pendant que 23 personnes autistes et 23 sujets contrôles (au comportement apparié) répondent à des questions -portant soit sur elles-mêmes, soit sur quelqu'un d'autre (la reine d'Angleterre!) - et sur des caractères soit physiques, soit mentaux.

Résultats principaux:

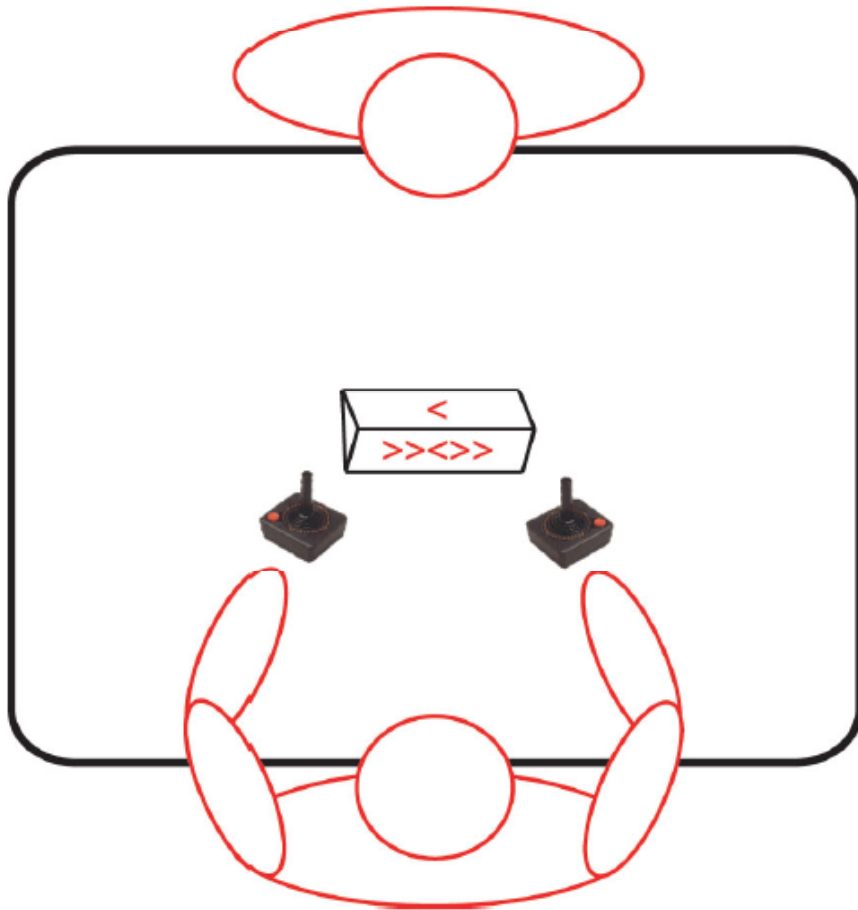
- deux régions qui répondent préférentiellement au « soi » chez le sujet normal (cortex cingulaire médian et cortex préfrontal ventromésial) ne le font plus chez les autistes.
- la différence soi/autre dans le cortex préfrontal ventromésial corrèle avec la sévérité des troubles du comportement social dans l'enfance.



Discussion de ces études d'imagerie du « soi »

- Les résultats convergent vers un réseau relativement reproductible du « soi », notamment au niveau du cortex préfrontal ventromésial
- L'interprétation des résultats reste ambiguë:
 - soit nous disposons d'une représentation détaillée de nous-mêmes, et nous utilisons ce réseau du « soi » pour **simuler** l'esprit des autres et tenter de le comprendre.
 - soit nous ne disposons pas d'un système spécifique d'introspection: notre connaissance est fondée sur l'observation de « soi-même comme un autre ».
- Les paradigmes expérimentaux sont peu rigoureux
 - nécessité absolue de développer de meilleurs protocoles métacognitifs, plus proches de la psychophysique
 - qui fassent notamment appel à l'introspection d'états mentaux réels et non à leur imagination par le biais d'histoires ou de photos.

Une même représentation cérébrale pour *mes* erreurs et *tes* erreurs



Dans des blocs différents, le participant **exécute** la tâche d'Eriksen, ou bien **observe** l'expérimentateur exécuter la tâche.

Exécution:

Le participant voit le stimulus central accompagné de distracteurs périphériques.

L'expérimentateur observe et, après chaque bloc, lui rapporte combien d'erreurs ont été faites.

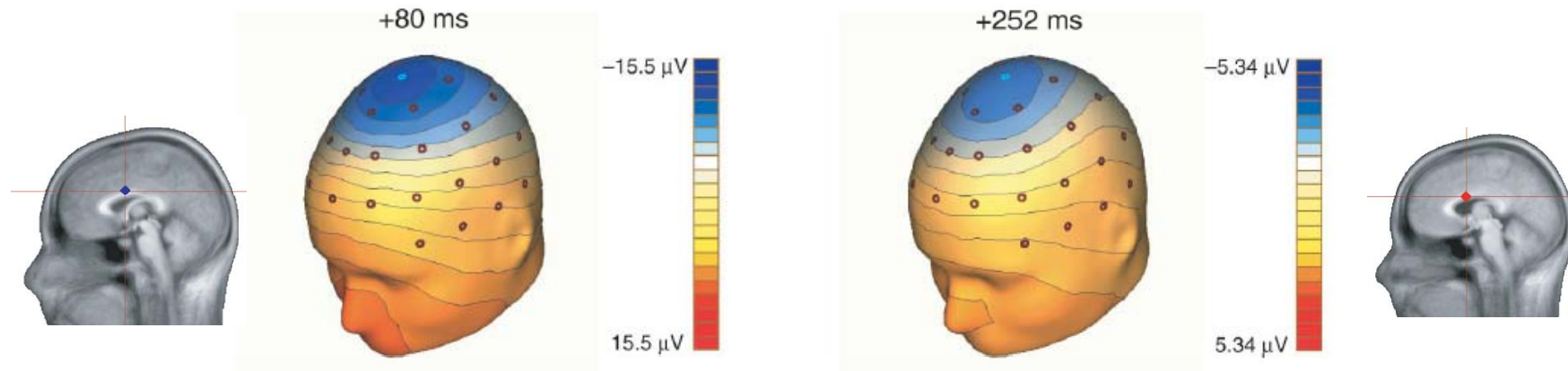
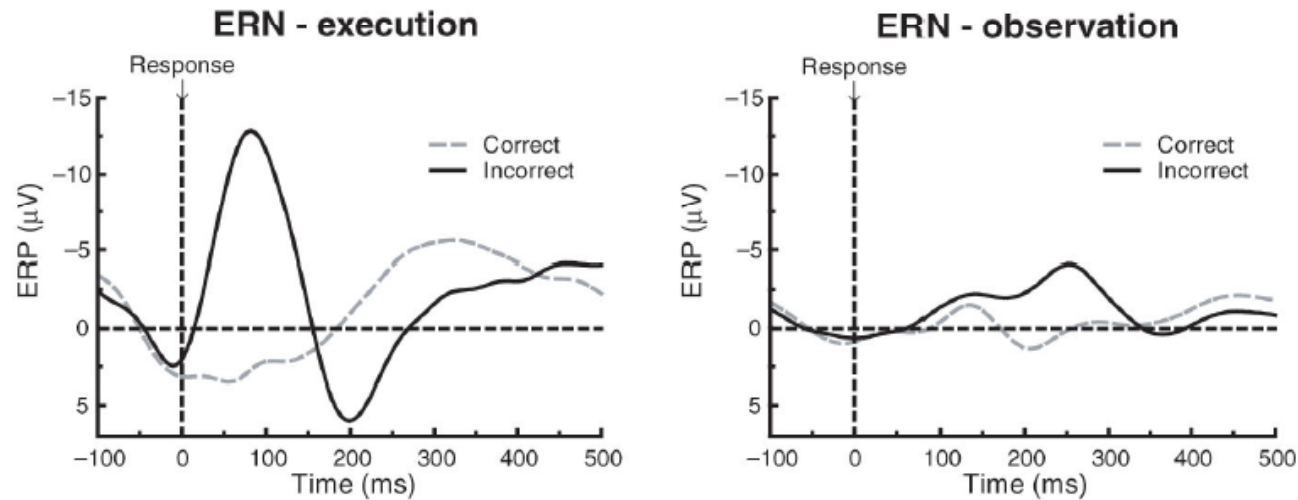
Observation:

Le participant ne voit que le stimulus central (ce qui lui permet de détecter plus facilement les erreurs).

Il fixe ce stimulus sans bouger les yeux, mais on lui demande de prêter attention aux réponses de l'autre.

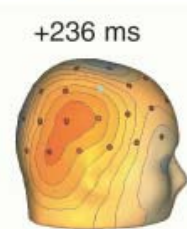
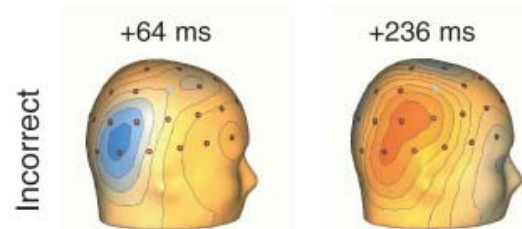
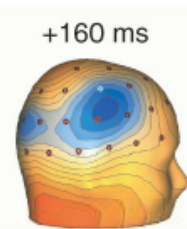
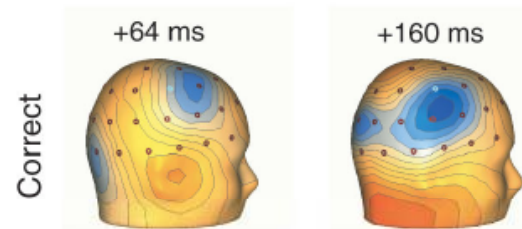
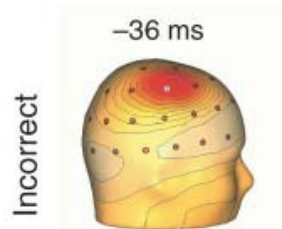
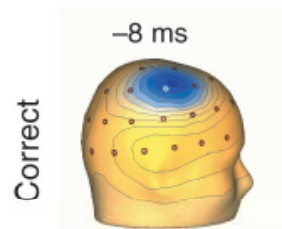
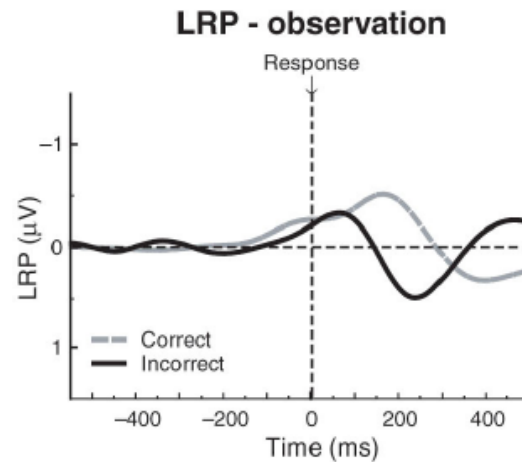
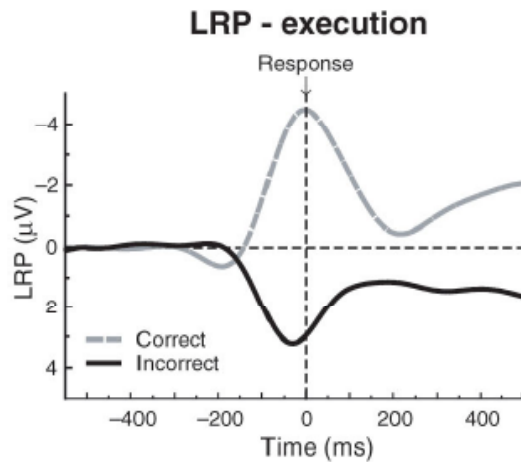
Les potentiels évoqués sont enregistrés, et utilisés pour calculer l'ERN (*error-related negativity*) et le LRP (*lateralized readiness potential*).

Une même représentation cérébrale pour *mes* erreurs et *tes* erreurs



van Schie, H. T., Mars, R. B., Coles, M. G., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nat Neurosci*, 7(5), 549-554.

Une même représentation cérébrale pour *mes* erreurs et *tes* erreurs



Le LRP, durant la période d'observation, montre deux étapes:

1. activation correcte du cortex moteur, qui *précède* la réponse de l'autre personne
2. activation du côté de la réponse faite par l'autre personne

Il y a donc simulation mentale:

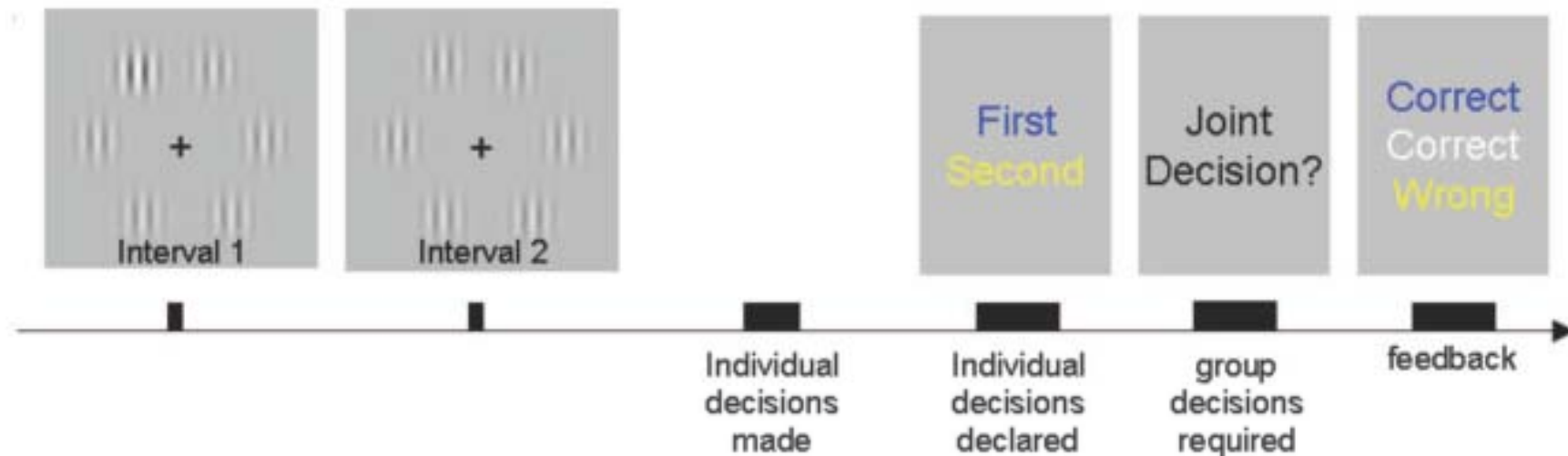
- au niveau moteur, de la réponse qui devrait être faite
- puis de celle qui est effectivement faite par autrui
- et enfin d'un signal d'erreur.

L'importance d'être conscient de ses propres capacités pour le comportement collectif

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010).
Optimally interacting minds. *Science*, 329(5995), 1081-1085.

Question posée: La prise de décision s'améliore-t-elle quand on demande à plusieurs personnes de se mettre d'accord?

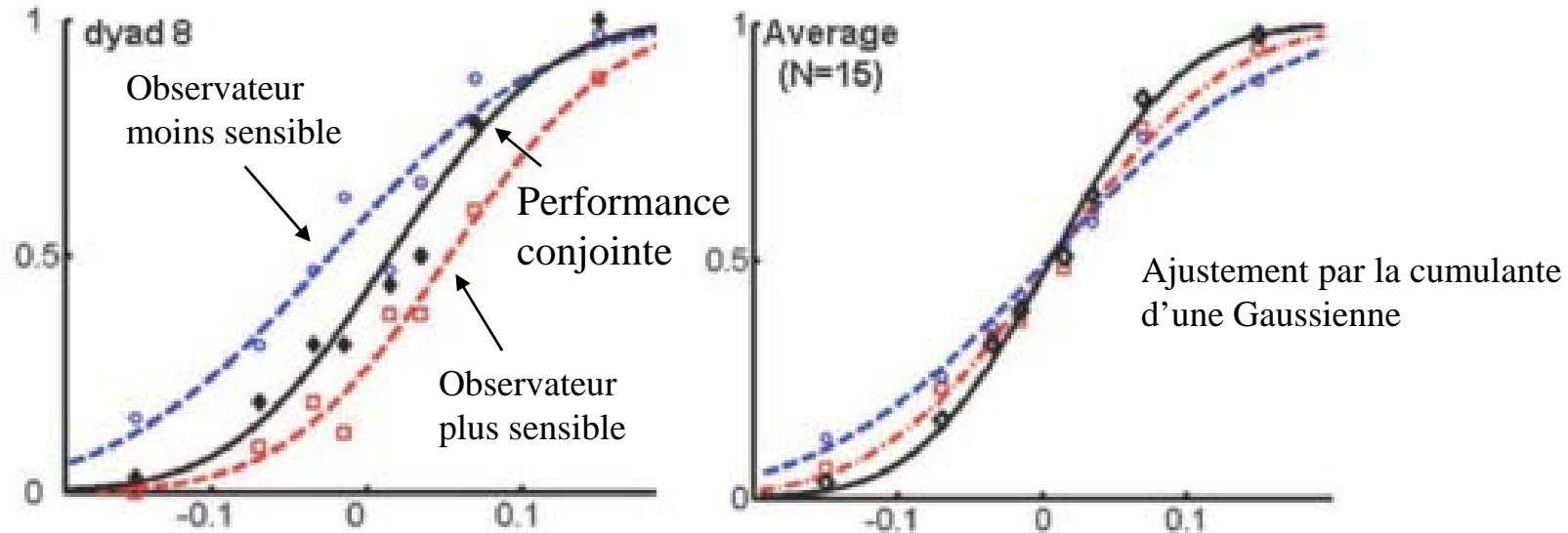
- Tâche psychophysique (décider quel intervalle contient une cible)
- Réalisée simultanément par deux sujets
- En cas de désaccord, l'expérience s'arrête et les deux personnes échangent jusqu'à ce qu'elles se mettent d'accord.



L'importance d'être conscient de ses propres capacités pour le comportement collectif

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010).
Optimally interacting minds. *Science*, 329(5995), 1081-1085.

La performance conjointe dépasse, en sensibilité (d-prime), celle de chacun des individus, de leur moyenne, ou même du meilleur des deux.



Elle ne peut s'expliquer qu'en supposant que les participants échangent leur **niveau de confiance** sur ce qu'ils ont vu à un essai donné (*weighted confidence sharing model*).

Des expériences ultérieures montrent (1) l'amélioration n'existe que si les sensibilités des deux participants ne diffèrent pas trop (2) la communication est indispensable, mais pas le feedback.

Pour Chris Frith, les capacités métacognitives et la conscience ont peut-être évolué précisément afin de faciliter l'échange entre individus et la prise de décision en société.

Conclusions

- Aux plus hauts niveaux de la méta-cognition consciente, la connaissance de soi et la connaissance de l'autre sont étroitement reliées:
 - elles se développent simultanément chez l'enfant
 - elles ne sont pas indépendantes, mais interagissent entre elles (de sorte que je confonds, en partie, ce que je sais et ce que tu sais)
 - elles font appel à des réseaux cérébraux en partie communs
- Ces observations restent vraies au niveau de la représentation de l'erreur, qui est un phénomène plus automatique
 - nous adoptons facilement la « posture intentionnelle » (*intentional stance*) et traitons les autres comme nous-mêmes.
- Sans doute l'introspection, fictive ou pas, joue-t-elle un rôle essentiel dans le dialogue social propre à l'espèce humaine.