

Introspection et métacognition :
Les mécanismes de la connaissance de soi

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

Cours

Liens entre conscience et métacognition

La métacognition est-elle possible en l'absence de conscience?

- La question paraît étrange:

Comment pourrions-nous avoir accès à des informations sur nous-mêmes, sans que cet accès soit automatiquement conscient?

Il nous semble, pratiquement par définition, qu'à chaque fois que nous « plongeons en nous-mêmes », c'est un acte d'introspection consciente.

Le modèle de l'espace de travail global implique que toute représentation consciente est rendue disponible à l'ensemble des traitements réflexifs.

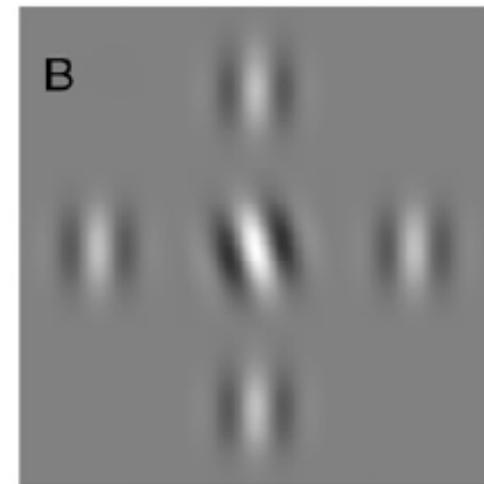
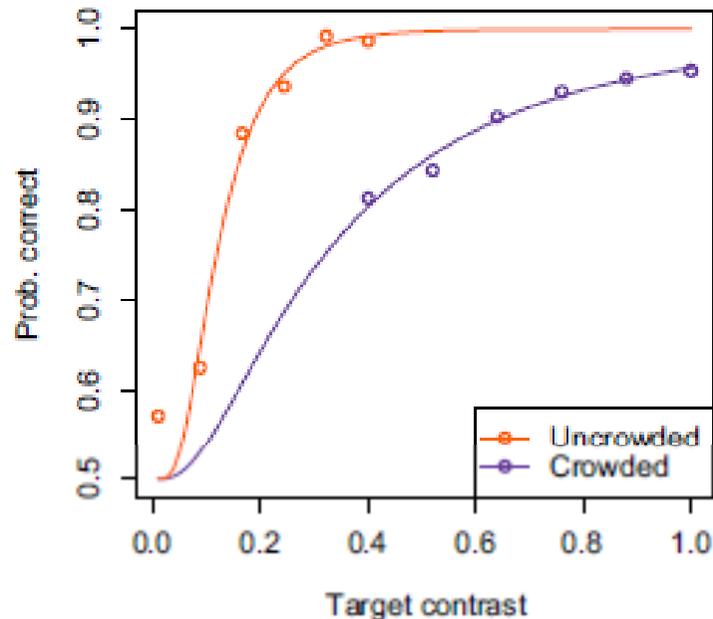
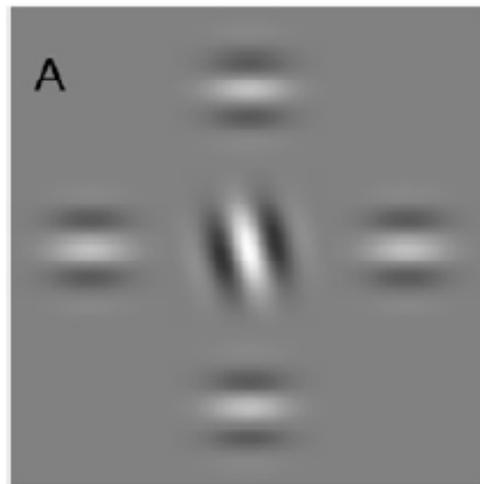
L'inverse est-il vrai? La théorie HOT 'Higher-order thought' de David Rosenthal (1997) implique que toute représentation intégrée à une « pensée d'ordre supérieur » est nécessairement consciente.

- Cependant, il n'y a rien dans la définition des processus métacognitifs qui interdise que les processus relativement élémentaires de « surveillance » des autres opérations mentales soient automatiques et non conscients.
- Je me propose de montrer que les jugements de confiance et d'erreur fournissent des exemples de métacognition sans conscience.
- Au passage, nous étudierons les mécanismes du jugement de confiance: comment savons-nous si nous sommes sur le point de nous tromper?

Le jugement de confiance en soi: Une compétence authentique

Barthelme, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proc Natl Acad Sci U S A*, 107(48), 20834-20839.

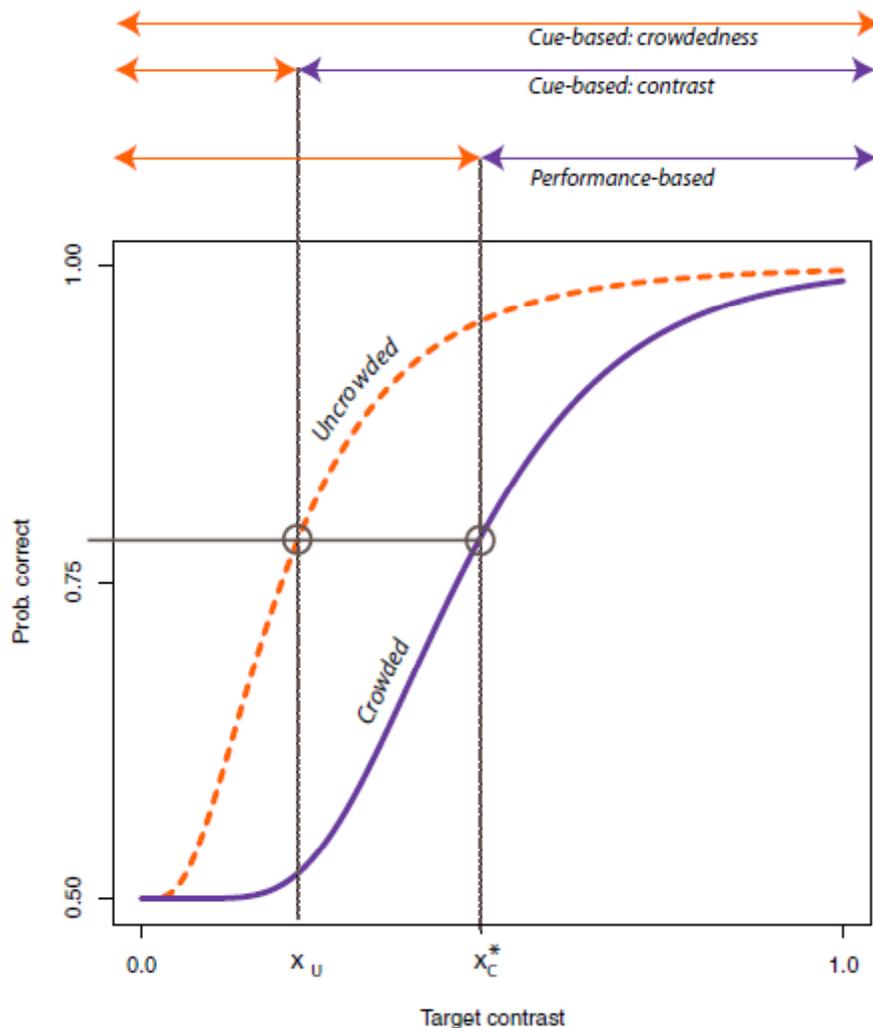
- On sait, au moins depuis Henmon (1911), que la performance et le jugement de confiance covarient souvent avec la difficulté de la tâche (exemple: lettres floutées)
- Barthelme et Mamassian (2010) notent que la confiance pourrait provenir, soit d'indices de bas niveau (le flou), soit d'une authentique évaluation de la probabilité de se tromper (représentation explicite de l'incertitude).
- Stimuli = 2 ensembles de Gabors en périphérie, variables = contraste et encombrement (*crowding*).
- Tâche = jugement de 2nd ordre: choisir le stimulus qui donne la plus grande confiance de répondre correctement à un jugement d'orientation; puis juger de son orientation



Le jugement de confiance en soi: Une compétence authentique

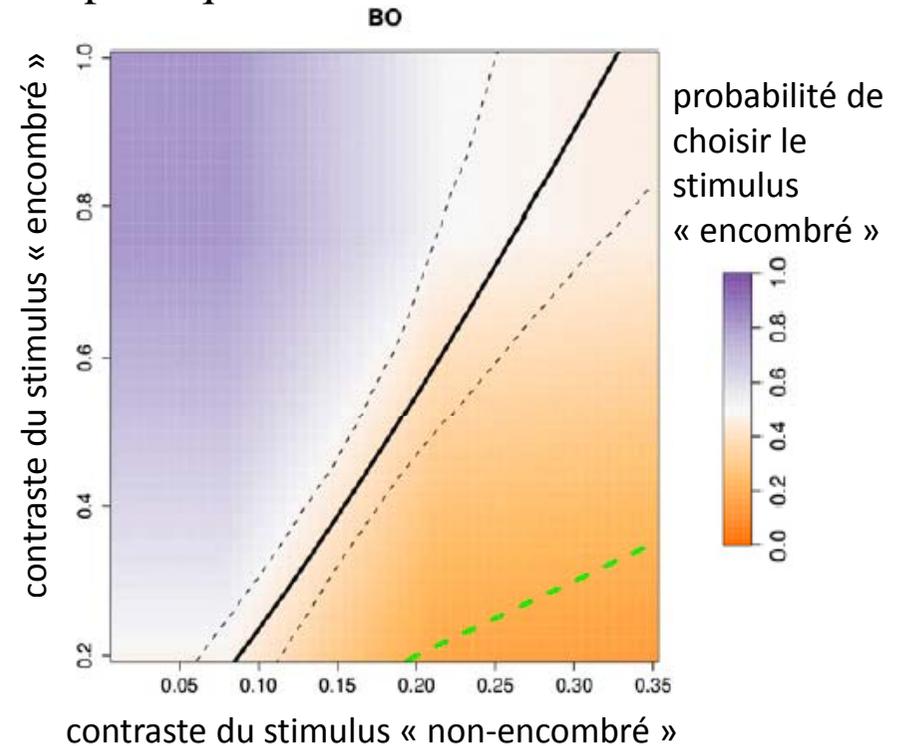
Barthelme, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proc Natl Acad Sci U S A*, 107(48), 20834-20839.

- Théorie: les performances devraient différer considérablement selon que les sujets utilisent un seul indice (contraste ou encombrement), ou qu'ils calculent l'incertitude réelle.



Résultats: tous les sujets utilisent un compromis entre les deux indices qui approxime le choix optimal.

Conclusion: ils disposent d'un mécanisme sophistiqué d'évaluation de l'incertitude



Une situation plus simple à théoriser: Le jugement de confiance en sa réponse ou « jugement de type II »

Décision de type I (tableau classique de la théorie de la détection du signal)

objective state of the world:	Respond "Absent"	Respond "Present"
Stimulus Present	Miss	Hit
Stimulus Absent	Correct Rejection	False Alarm

Décision de type II (*second-order signal detection theory*)

objective state of the world:	Respond "Error"	Respond "correct"
Correct response	2 nd order Miss	2 nd order Hit
Error	2 nd order Correct Rejection	2 nd order False Alarm

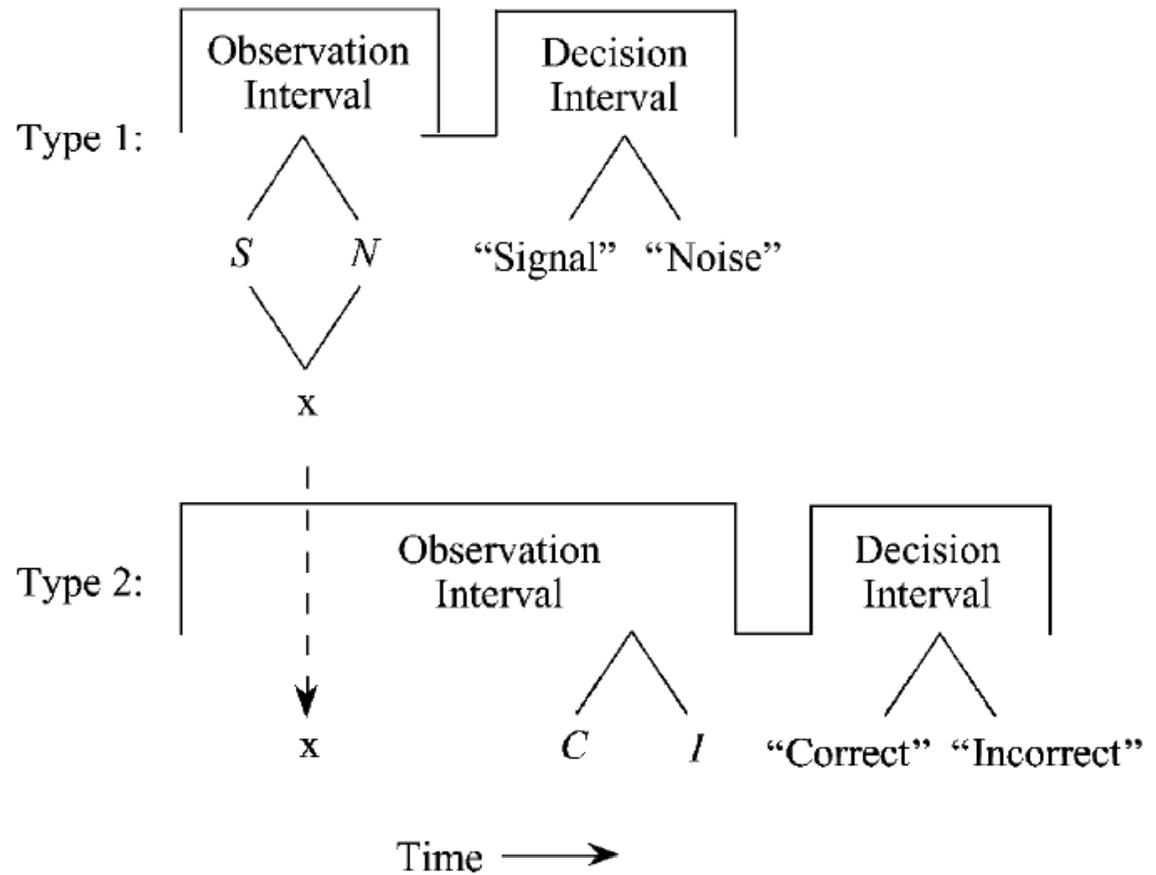
Au lieu d'une réponse binaire, il est également possible de remplacer la réponse II par un continuum de « confiance dans la réponse », sur une échelle numérique.

Comment modéliser le jugement de confiance en sa propre réponse?

Le modèle du continuum unique (*type II signal detection theory*)

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon Bull Rev*, 10(4), 843-876.

- Ce modèle part de la théorie de la détection du signal: l'observateur recueille des observations sur un continuum bruité et fonde sa décision (de type I) sur l'application d'un critère précis.
- La décision de type II est fondée sur une estimation de la probabilité de se tromper, *calculée à partir des mêmes données*



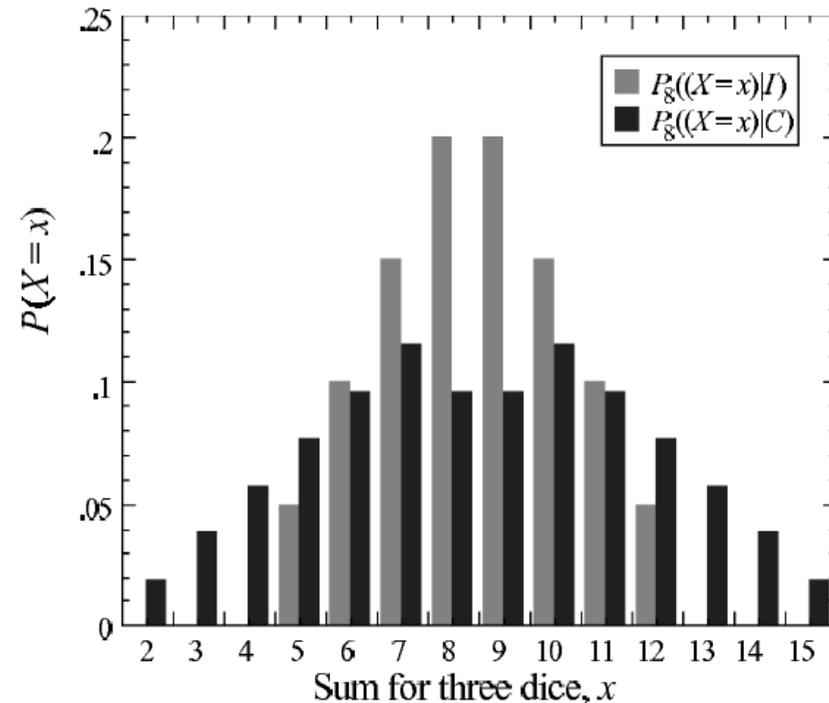
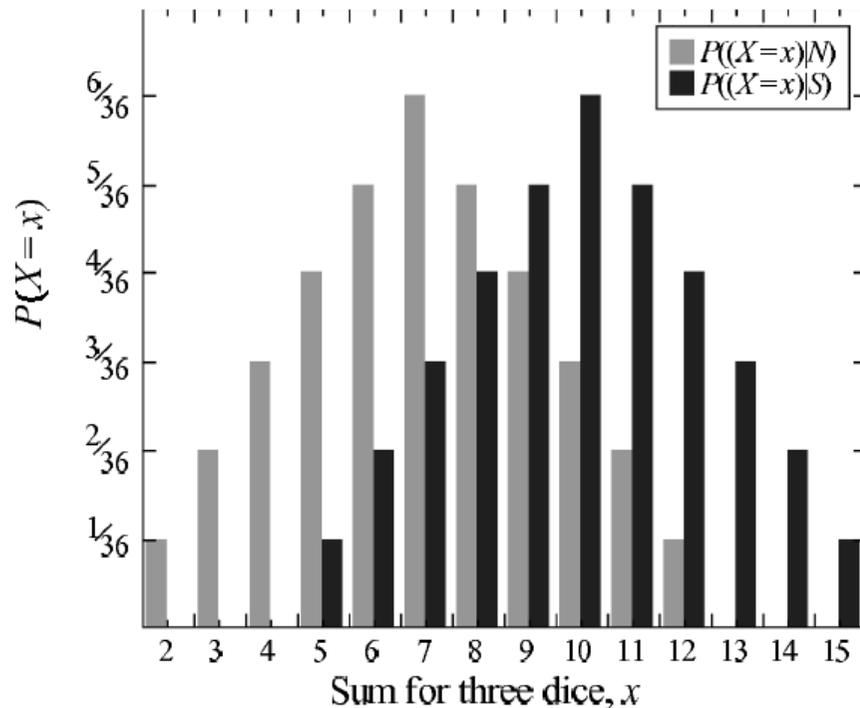
Comment modéliser le jugement de confiance en sa propre réponse?

Le modèle du continuum unique (*type II signal detection theory*)

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon Bull Rev*, 10(4), 843-876.

- Exemple: lancer de trois dés, dont deux sont normaux, l'un a trois faces « 3 », et trois faces « 0 ».
- Observation = somme des trois dés.
- Tâche = deviner la valeur du dé spécial.
- Méthode optimale (Green & Svets) = application d'un critère fixe à la somme

- Pour un critère donné (par exemple $\text{somme} > 8$), on peut calculer les *distributions de probabilité de type II*: probabilité d'observer une certaine somme lors des essais corrects (resp. incorrects).

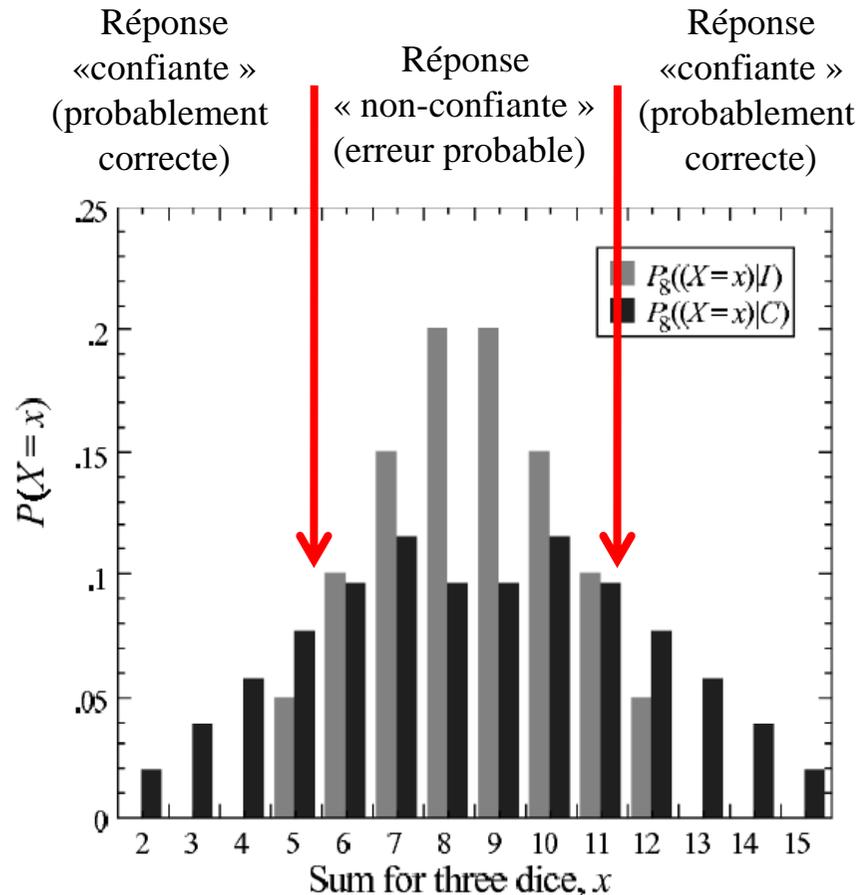


Comment modéliser le jugement de confiance en sa propre réponse?

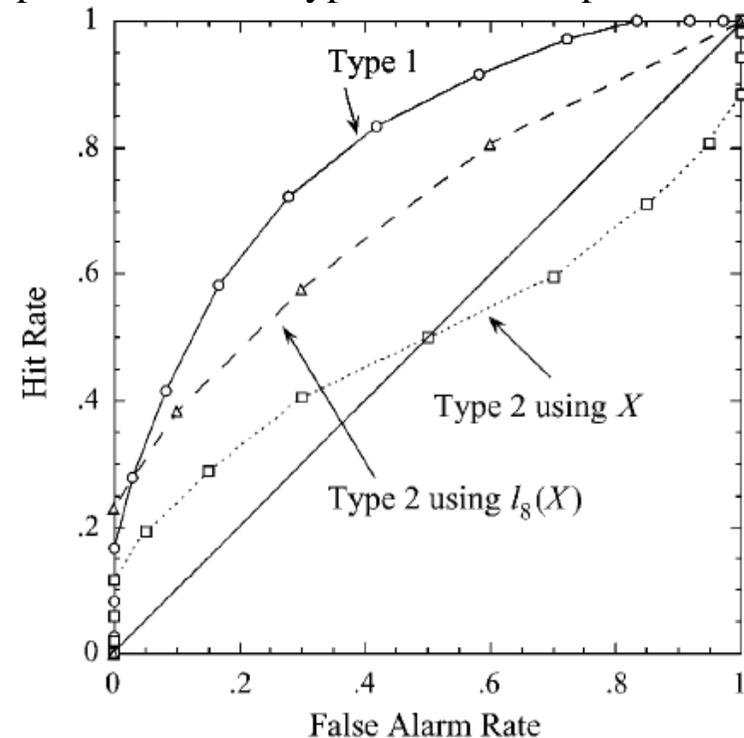
Le modèle du continuum unique (*type II signal detection theory*)

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon Bull Rev*, 10(4), 843-876.

- Pour une fonction de coût donné, la décision optimale consiste à appliquer un critère fixe au ratio de ces distributions.



- En déplaçant ce critère, on obtient alors une « courbe ROC » de type II
- Noter que celle-ci dépend des *deux* critères (de type I et de type II). Il existe donc une famille de courbes ROC de type II
- La performance de type II est donc dépendante de la performance de type I, et bornée par celle-ci.



Importance et limites de la théorie de Galvin et al.

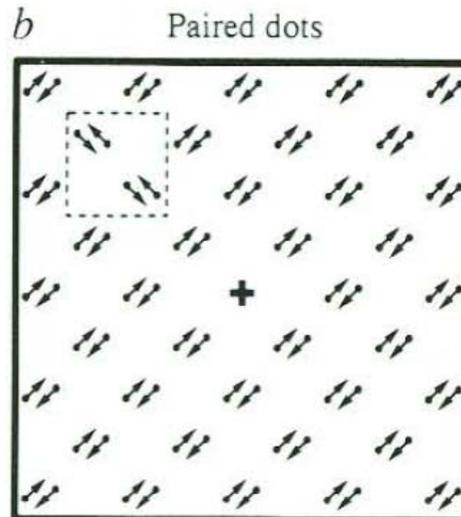
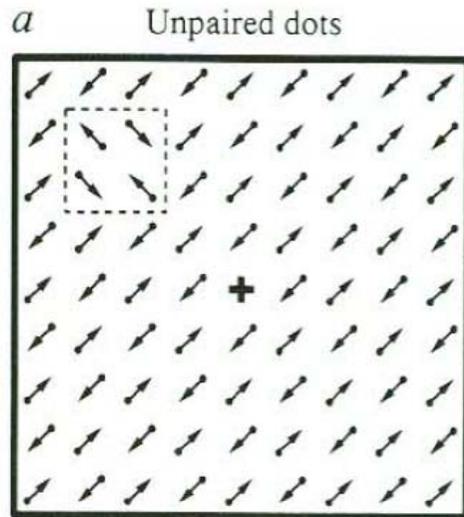
- Dès qu'un « signal » permet une détection de premier ordre meilleure que le hasard, il permet également un jugement de confiance de second ordre meilleur que le hasard.
- Ce niveau de « réussite de type II » est calculable (voir le concept de meta-d' et le programme Matlab fourni par Rounis, Maniscalco, Rothwell, Passingham & Lau, *Cognitive Neuroscience, sous presse*). La théorie montre que la performance de type II (mesurée par l'aire sous la courbe) est toujours moindre que la performance de type I.
- De façon plus générale, la perspective Bayésienne implique que chaque aire cérébrale estime non seulement la représentation la plus probable, mais également sa distribution de probabilité, ce qui fournit un code local de l'incertitude.
- Prédications (sous l'hypothèse que le même signal est utilisé pour les décisions I et II):
 - Le jugement de confiance devrait toujours être meilleur que le hasard quand la performance de type I l'est.
FAUX!
 - Le jugement de confiance peut être fondé sur un calcul modulaire, donc potentiellement non conscient.
VRAI!

Notre conclusion sera différente:

il existe de multiples origines pour le jugement de confiance (plusieurs systèmes de détection d'erreur ou de jugement métacognitif), certains conscients, d'autre non.

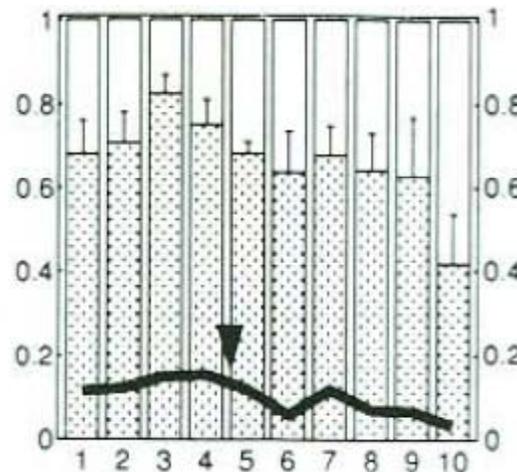
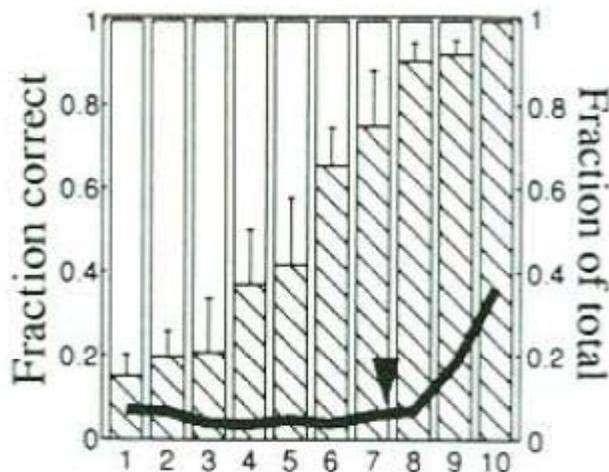
La vision aveugle: Une dissociation entre performance et confiance?

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature*, 377(6547), 336-338.



Le degré de confiance corrèle avec la réussite objective

Aucune corrélation!



Stimuli = textures de points en mouvements

Deux conditions:

- l'une **visible**: une région se distingue par l'utilisation de points en mouvement perpendiculaire à ceux du fond

- l'autre **invisible**: les points sont appariés pour qu'ils aient des mouvements opposés.

Tâche = décider où se trouve la cible (1 position parmi 4 possibles) puis jugement de confiance (1 à 10)

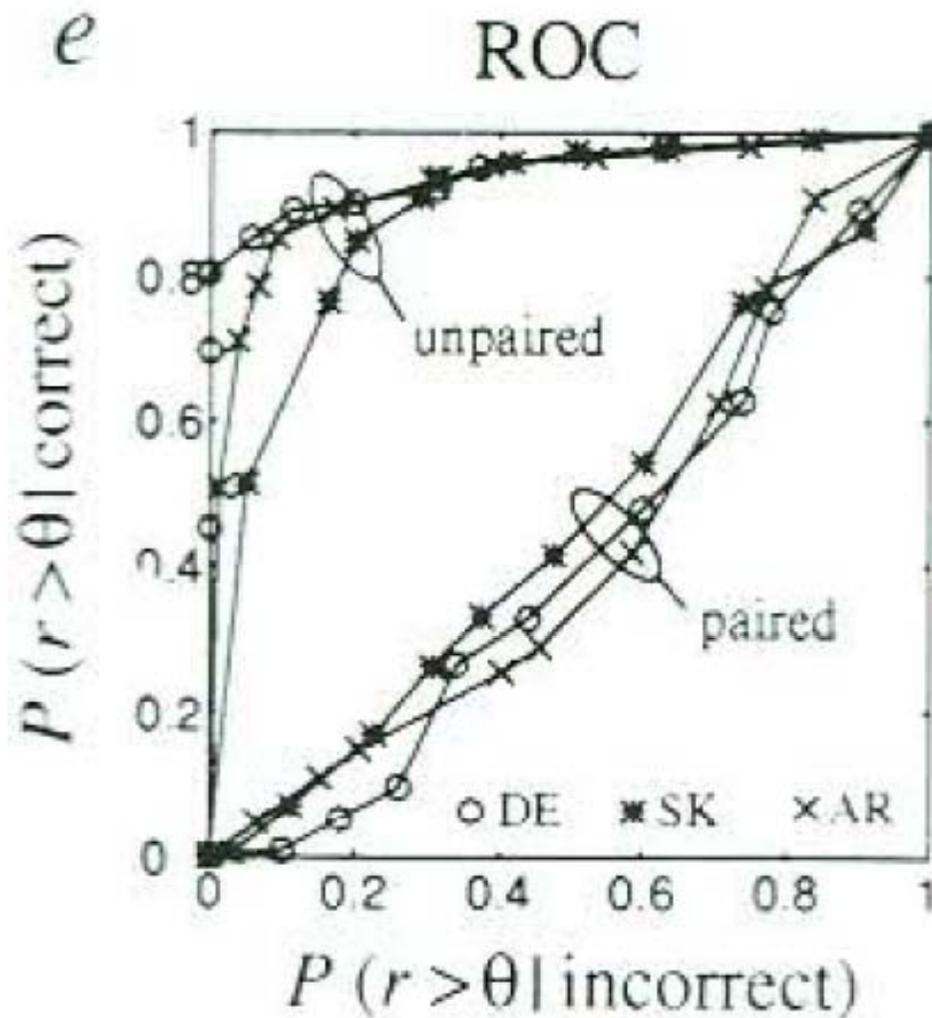
Résultats:

- performances équivalentes et très supérieures au hasard dans les deux cas (74.4 versus 69.9% correct).

- mais jugement de confiance impossible dans le cas invisible.

La vision aveugle: Une dissociation entre performance et confiance?

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature*, 377(6547), 336-338.



Une autre manière de présenter les résultats:
La courbe ROC (*receiver operating characteristic*) de type II.

Pour chaque niveau de confiance, on indique la proportion d'essais corrects et incorrects ayant atteint ce niveau.

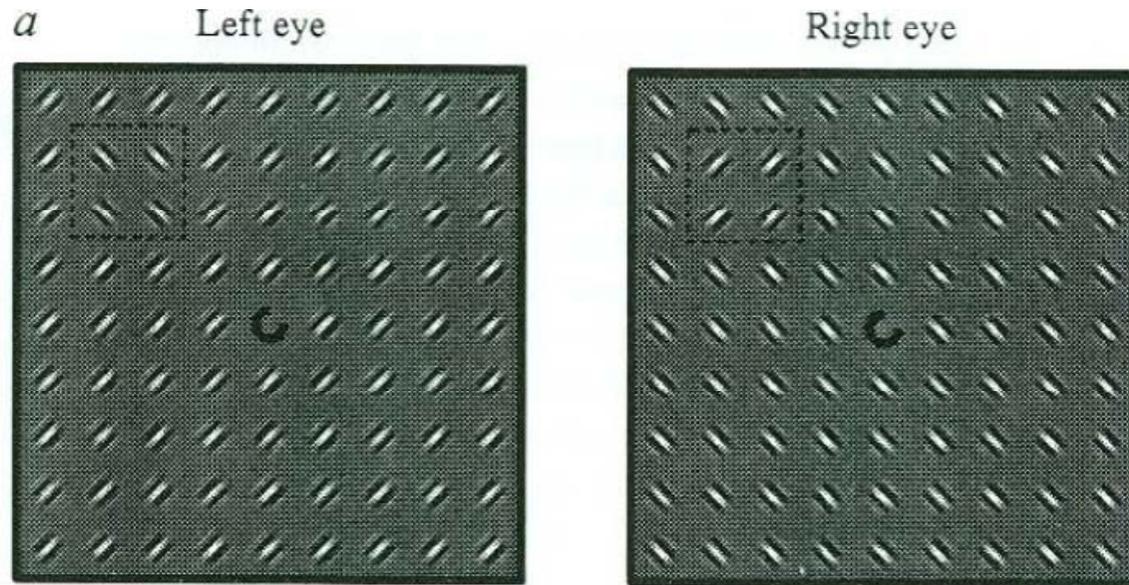
On voit que, pour les essais visibles (*unpaired*), dès que la confiance monte, les essais corrects excèdent en nombre les essais incorrects.

Ce n'est pas le cas pour les invisibles (*paired*).

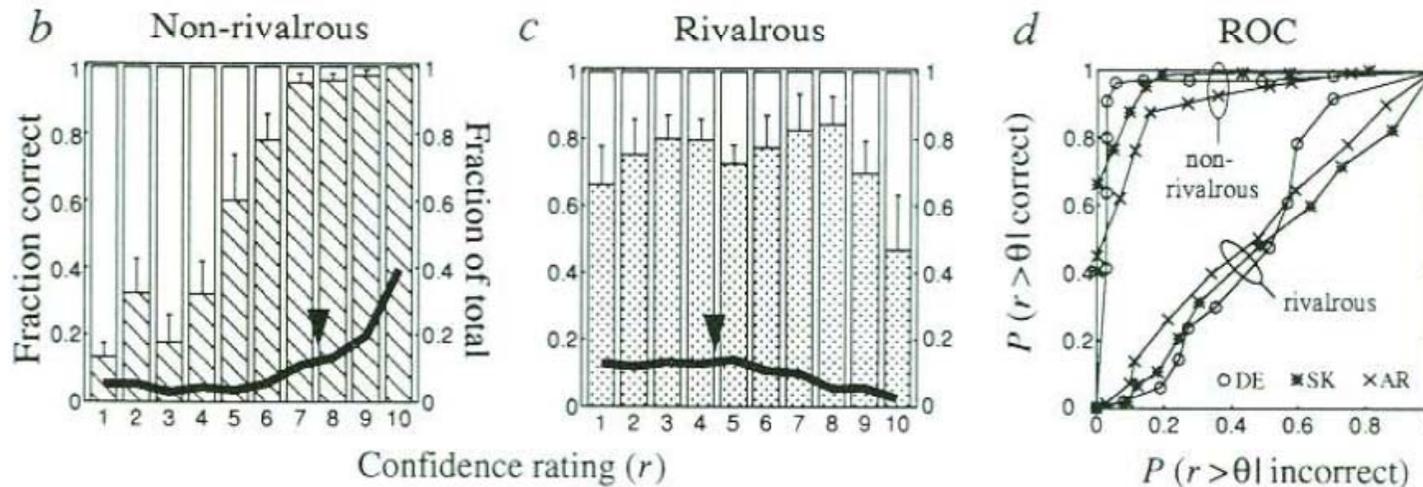
Il s'agirait donc d'une forme de vision aveugle (*blindsight*) chez le sujet sain.

Pour des résultats similaires chez le patient GY, voir Persaud et al., *Nature Neuroscience* 2007 (cours de 2009)

Vision aveugle – Dissociation complète entre performance et confiance



Réplication de ces résultats dans une situation de rivalité binoculaire



Non-réplication

Morgan et al (Nature, 1997)

« Clearly the experience, although vague, is not unconscious »

« Kolb and Braun's subjects were instructed to use the full confidence scale, irrespective of their absolute sense of certainty. If they were reluctant to use the vague cues in the masked condition as evidence for a high-confidence judgment, they may have decided to produce ratings randomly, with the result that Kolb and Braun observed.

It is otherwise difficult to understand why they made errors in trials when they claimed to be highly confident: the opposite of blindsight. »

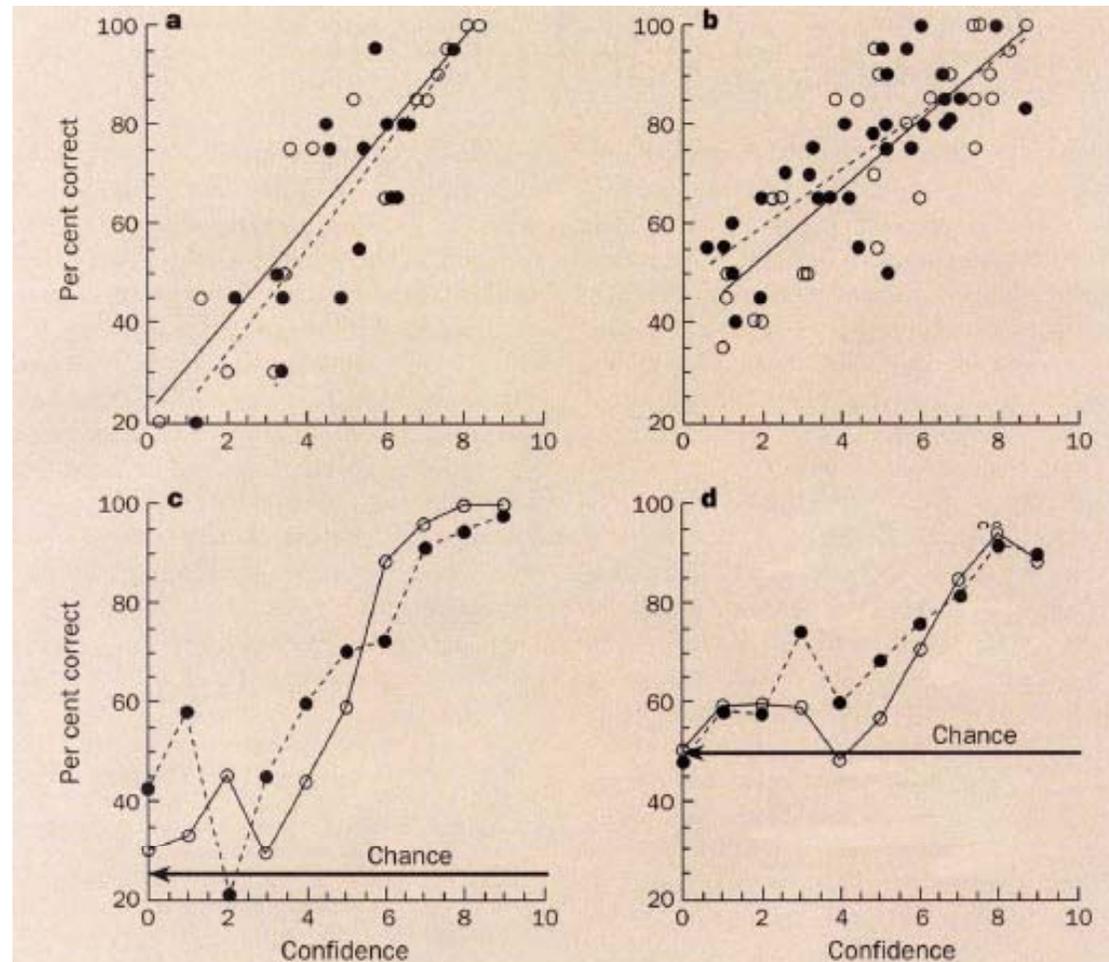


Figure 1 Correlation between success and confidence of success. Each point in a and b represents the performance of an individual subject at a particular exposure duration, in either the masked (solid symbols) or the unmasked (open symbols) condition. Linear regressions are shown (masked condition, broken lines; unmasked condition, solid lines). All correlation coefficients are high (a, 0.84, 0.92; b, 0.77, 0.92; masked, unmasked). In experiment 1 (a) the subjects were the three authors and masked and unmasked trials were randomly interleaved. In experiment 2 (b) the two conditions were presented separately and the subjects included three naive observers. Further observations were made of one other experienced subject (D. I. A. M.) in the masked condition alone (see Table 1). Details of the methods are available upon request from M. J. M. c, d, Replotting of the data from a and b, respectively. Each point represents the mean for all of the subjects, collapsed across exposure duration. Irrespective of duration, confidence is still a good predictor of success rate in both unmasked and masked conditions.

Le degré de confiance peut-il être dissocié de la performance?

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Conscious Cogn*, 10, 294-340.

- Kunimoto et al. proposent d'utiliser le jugement de type II, sous forme de pari financier, comme une mesure de la conscience
 - A chaque essai, le sujet répond d'abord à la tâche I (déterminer lequel de 4 mots est présenté masqué), puis parie sur sa réponse avec un jeton vert (risqué) ou rouge (peu risqué)
- (anticipation complète de l'article de Persaud et al., *Nature Neuroscience*, 2007 – qui ne le cite pourtant pas!)
- Kunimoto et al. montrent une dissociation faible mais significative entre les performances de type I (tâche directe) et de type II (jugement métacognitif de confiance).

Réponse objective	Jeton rouge	Jeton vert
Réponse correcte	+ 0.03 \$	+ 0.05 \$
Erreur	- 0.01 \$	- 0.03 \$

Le degré de confiance peut-il être dissocié de la performance?

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Conscious Cogn*, 10, 294-340.

- Expérience 1: Comparaison de la méthode de pari avec la méthode de Cheeseman-Merikle : estimer sa performance après chaque bloc
- Résultats: la méthode du pari s'avère bien plus sensible à de faibles améliorations de la performance

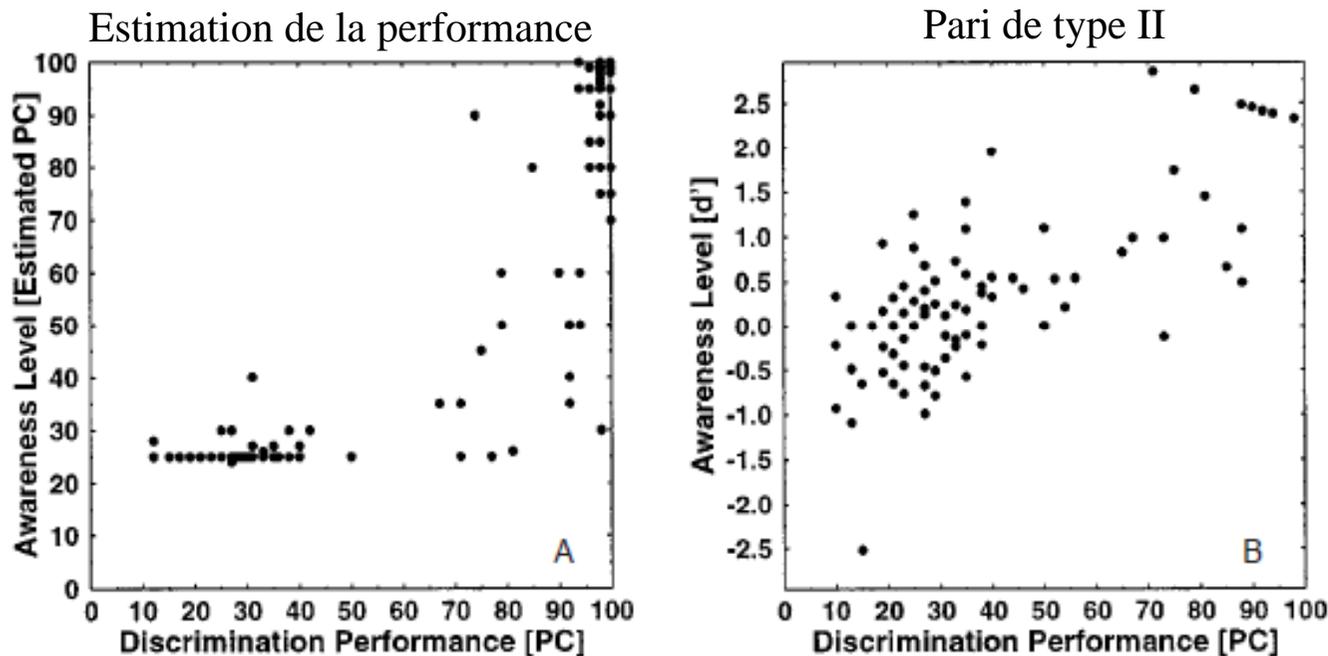


FIG. 3. Awareness as a function of discrimination performance for the (A) C&M procedure and (B) SDT procedure. (Each point represents one block of trials. The first three blocks for each subject are not plotted; blocks in which subjects' discrimination performance was 100% are also not plotted.)

Kunimoto et al. en concluent que c'est une mesure plus sensible de la perception consciente des mots masqués. Mais ce raisonnement est circulaire, on pourrait tout aussi bien en conclure que l'évaluation de la confiance est, au moins en partie, non consciente.

Le degré de confiance peut-il être dissocié de la performance?

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Conscious Cogn*, 10, 294-340.

- Expériences 2, 3 et 4: Mesure fine du seuil de masquage
 - d'une part, pour la performance sur la tâche I
 - d'autre part, pour la performance de la méta-tâche II

Résultats: le seuil de masquage est un tout petit peu plus bas pour la tâche de type I que pour la tâche de type II (19 ms versus 23 ms!)

Cela signifie qu'il existe effectivement des stimuli qui dissocient la performance de type I (meilleure que le hasard) et le jugement de confiance (nul).

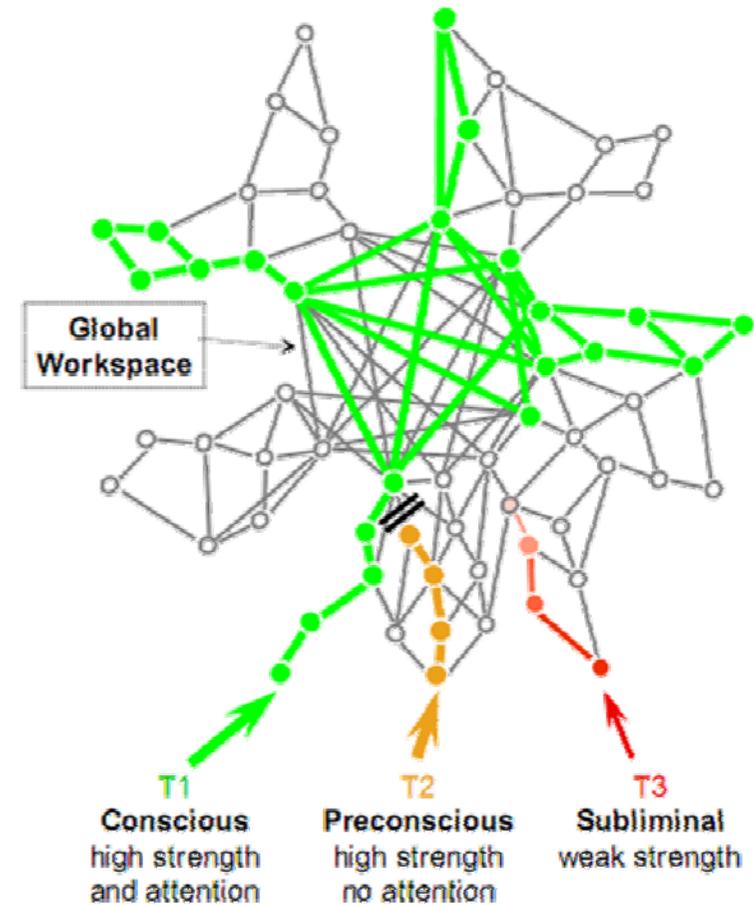
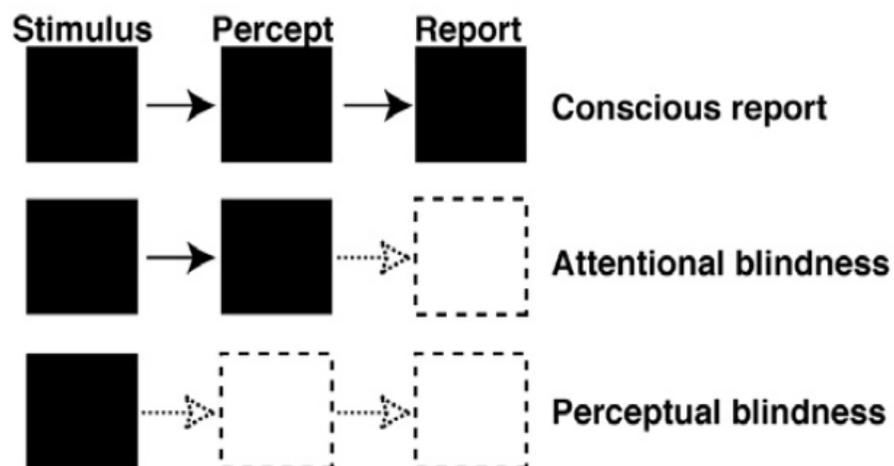
Ces résultats sont confirmés dans l'article de Persaud et al (*Nature Neuroscience* 2007) dans différentes conditions: le *blindsight*, l'apprentissage de grammaires, la tâche de casino de l'Iowa.

Toutefois la différence est souvent faible, et le point crucial n'est jamais testé – la confiance peut-elle être meilleure que le hasard lors des essais où les sujets rapportent n'avoir rien vu?

Dissociation entre la visibilité, la confiance et la performance

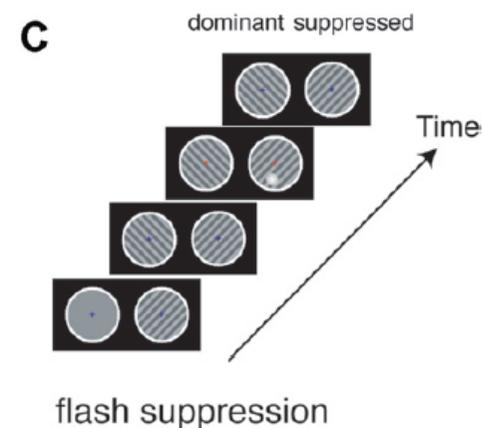
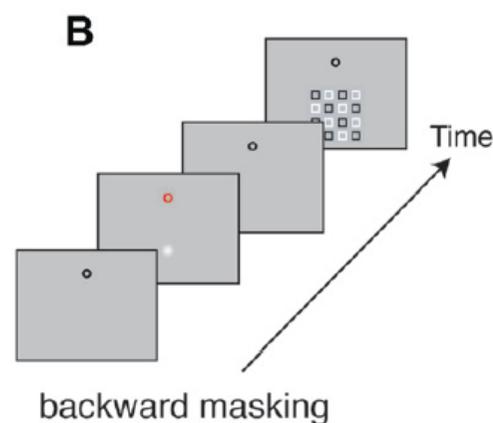
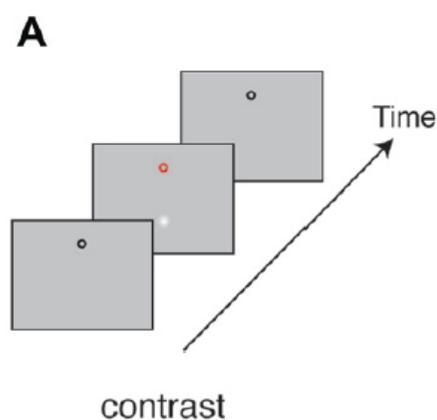
Kanai, R., Walsh, V., & Tseng, C. H. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Conscious Cogn.*

- Quel est le lien entre visibilité et confiance?
 - Selon Kanai et al., ce lien pourrait dépendre du facteur clé à l'origine de l'invisibilité: limitation perceptive ou limitation attentionnelle
 - Dans ce dernier cas, les essais « non-vus » proviennent d'une inattention temporaire. Les observateurs devraient être capables de juger, a posteriori, s'ils étaient attentifs ou pas. On devrait donc observer une corrélation entre leur confiance et leur performance.

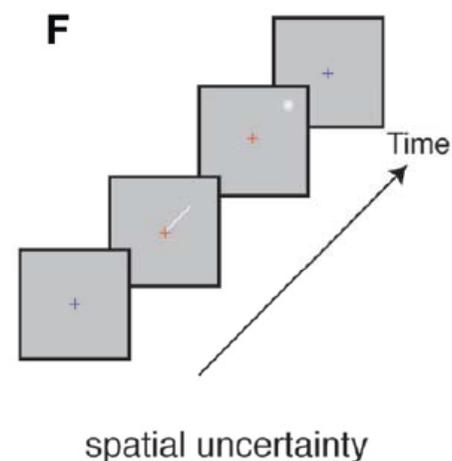
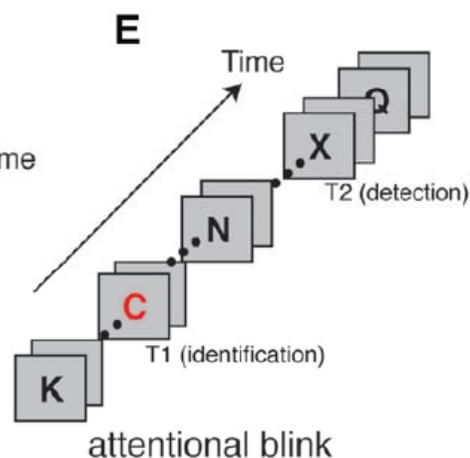
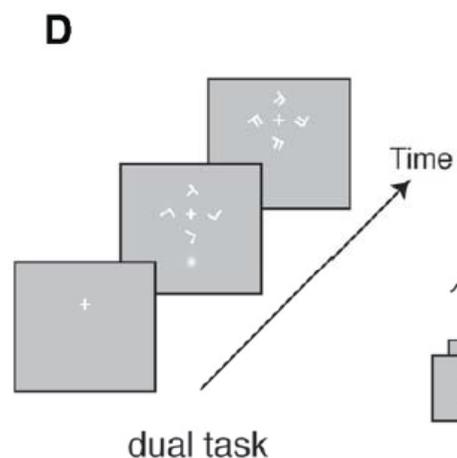


6 paradigmes expérimentaux

Cécité
d'origine
perceptive



Cécité
d'origine
attentionnelle



- A chaque essai, les participants rapportent
 - la présence ou l'absence d'un stimulus
 - le degré de confiance dans leur réponse: bas, moyen, haut

Dissociation entre la visibilité, la confiance et la performance

Kanai, R., Walsh, V., & Tseng, C. H. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Conscious Cogn.*

- Introduction d'un nouvel index de confiance: « *subjective discrimination of invisibility* » (SDI):
 - Très semblable au calcul du d-prime de type 2
 - Mais fondé uniquement sur les essais où le participant rapporte l'absence du stimulus
 - Logique: examiner uniquement les essais rapportés comme « non-vus », afin d'examiner l'origine de l'invisibilité

Jugement de type II:
Confiance subjective dans la réponse

Jugement de type I:
la réponse est
objectivement Correcte

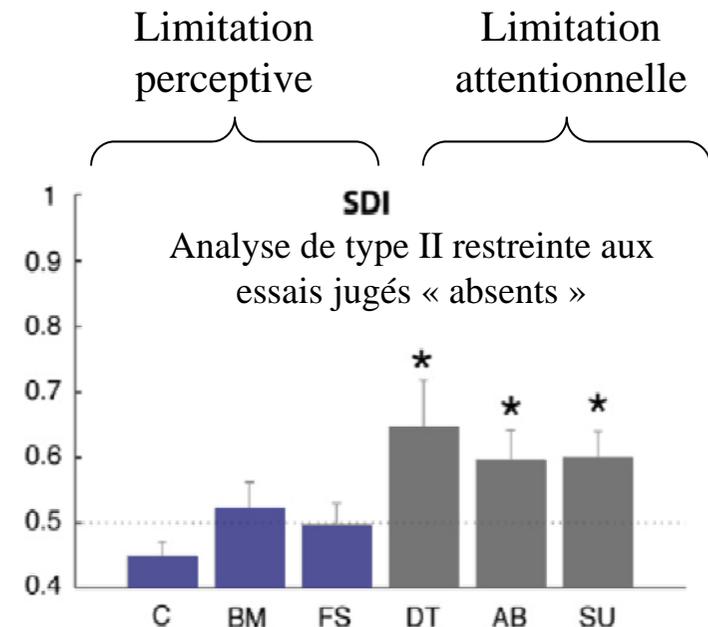
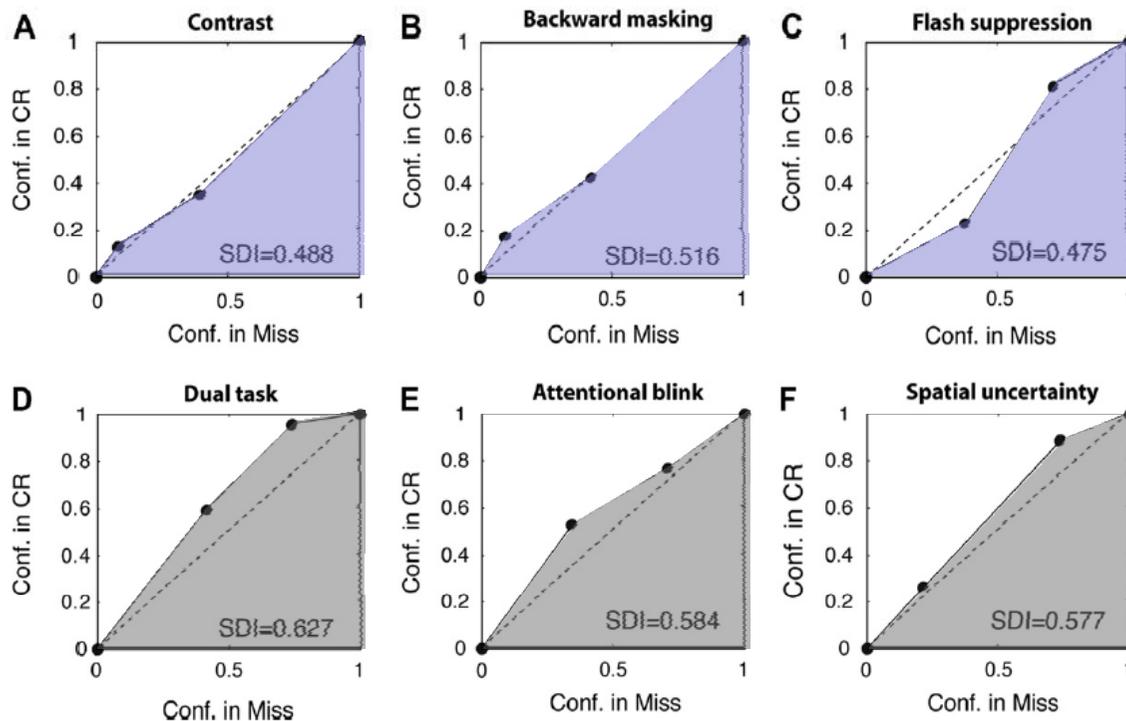
	High Confidence	Low Confidence
Corr. Reject	Type II Hit	Type II Miss
Erronée	Type II False Alarm	Type II Corr. Reject.

Résultats: La confiance corrèle avec la performance uniquement lors des essais où l'invisibilité est d'origine attentionnelle

La performance de type I est toujours meilleure que le hasard.

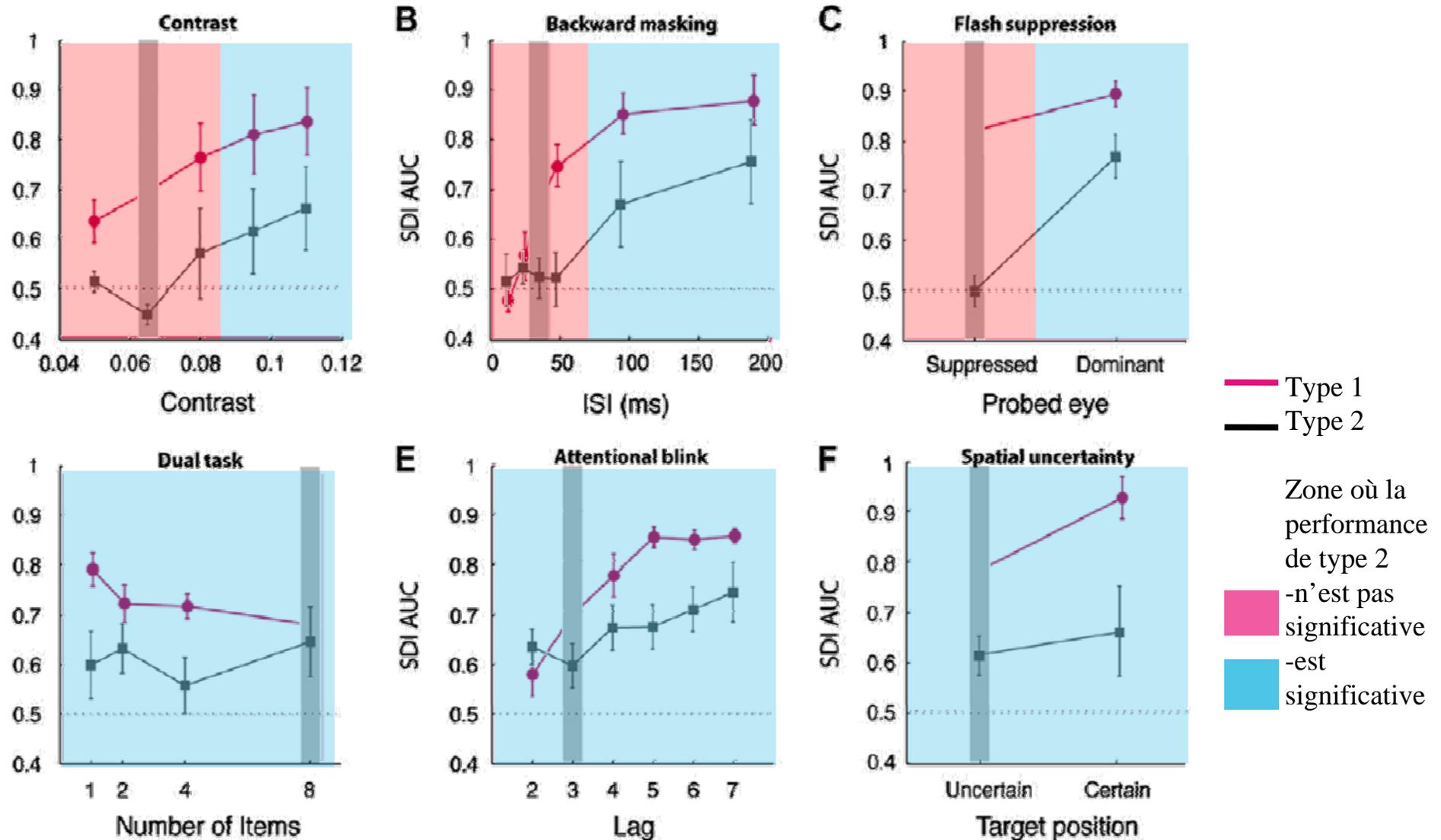
Analyse de type « ROC » (*receiver operating characteristic*) pour la performance de type II:

- Elle est meilleure que le hasard si la confiance est plus élevée pour les essais corrects que pour les essais erronés, c'est-à-dire que la courbe s'écarte de la diagonale



La dissociation entre confiance et performance n'existe qu'à faible niveau de visibilité

Même pour les trois premières conditions, lorsque la visibilité augmente, réapparaît une corrélation entre confiance et performance de type 1.

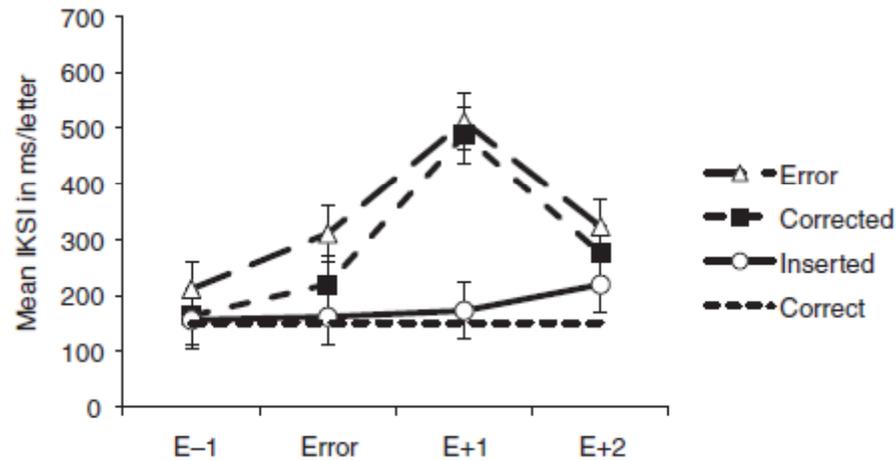


Conclusions de cette étude

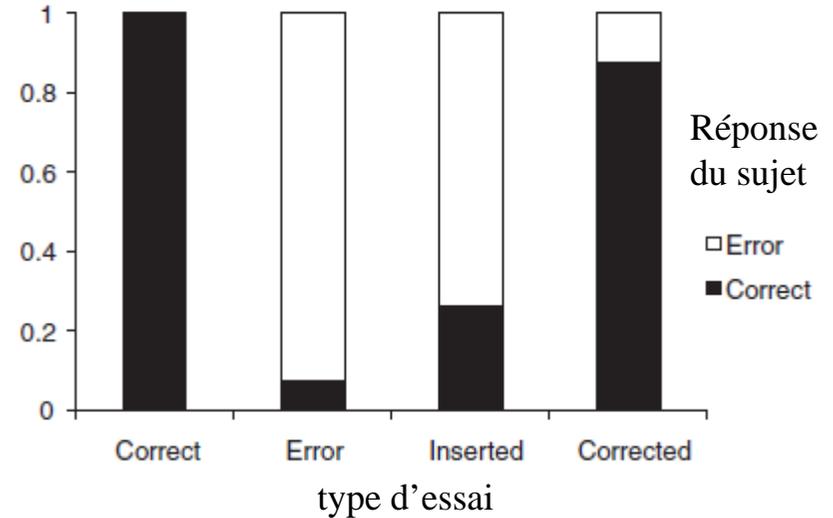
- L'expérience de Kanai et al. confirme à nouveau l'existence de conditions (masquage, *flash suppression*) dans lesquelles la performance de type I est supérieure au hasard, et la confiance (performance de type II) est nulle.
 - Ces résultats ne peuvent pas s'expliquer dans la théorie de Galvin et al , selon laquelle la confiance est issue des mêmes données et des mêmes processus que la décision initiale.
- Kanai et al. sont les premiers à dissocier les trois paramètres: réponse de type I, réponse de type II, et visibilité
- Leurs résultats montrent que le jugement de confiance peut être meilleur que le hasard, alors même que le sujet juge n'avoir rien vu. Dans les conditions d'invisibilité attentionnelle, les participants modulent correctement leur confiance en fonction de leur performance, alors même qu'ils n'ont pas perçu le stimulus.
- *Ergo*, le jugement de confiance n'est pas identique au jugement de visibilité du stimulus. Entrent probablement en considération des facteurs tels que la quantité d'attention et de vigilance avant même la présentation du stimulus.
- Les données restent compatibles avec l'idée que la confiance résulte d'un calcul *conscient* (mais portant, non seulement sur la perception du stimulus, mais également sur d'autres indices tels que la conscience de la position spatiale de l'attention, la conscience de son état de vigilance, etc...)

De multiples niveaux de détection d'erreurs

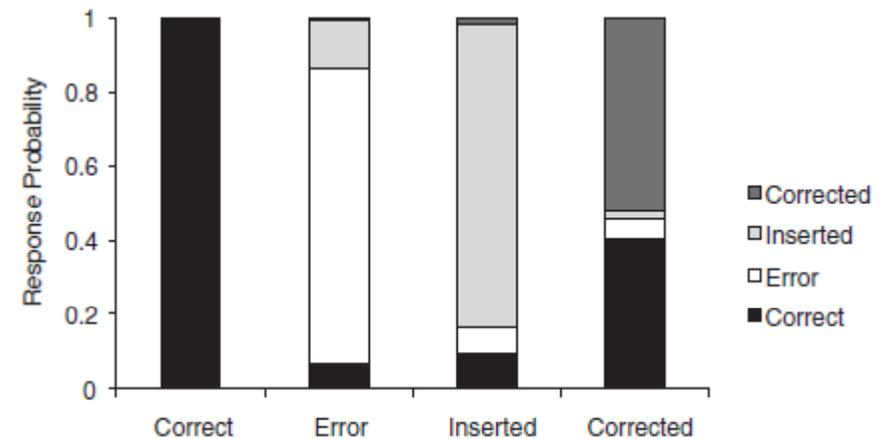
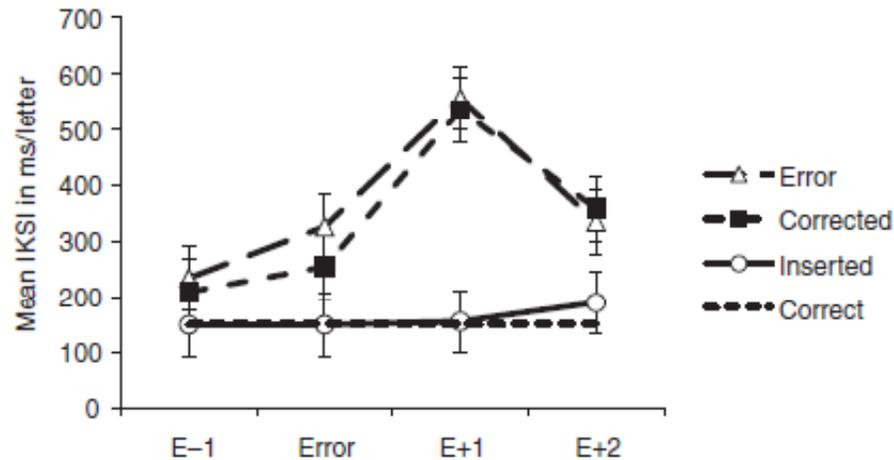
Le ralentissement après une erreur ne dépend que la performance réelle, pas de la rétroaction donnée au sujet



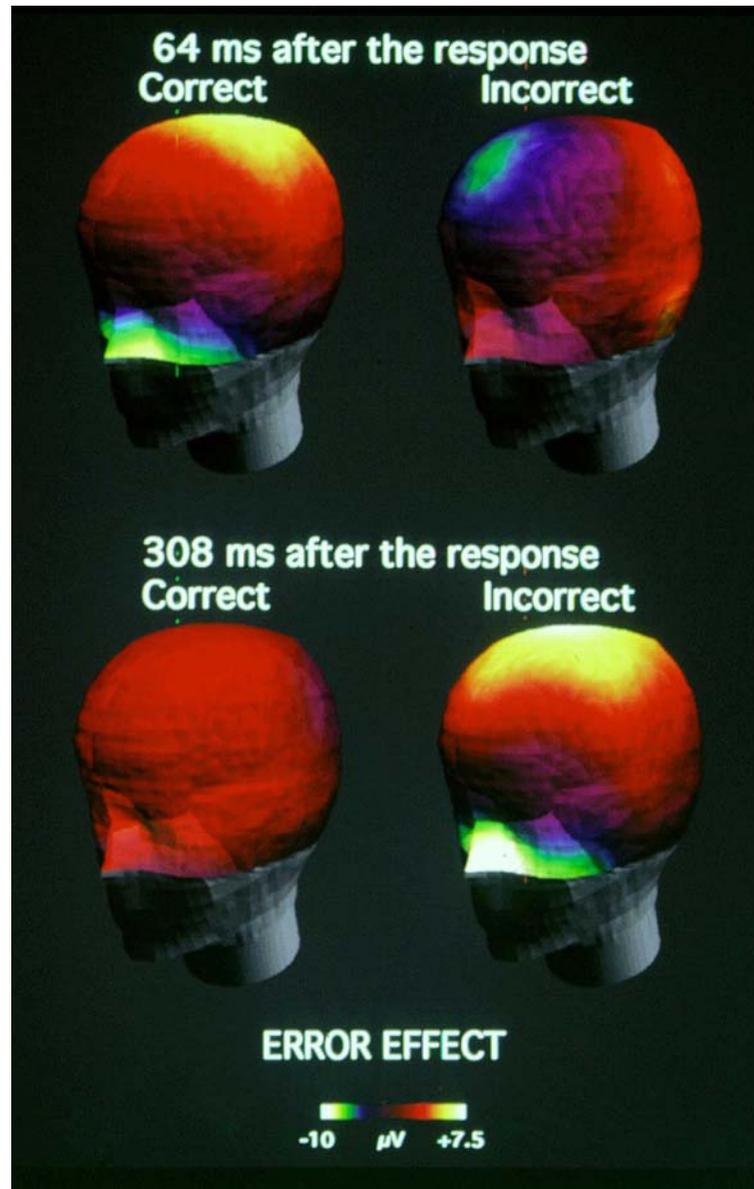
Consciemment, après chaque essai, les sujets détectent peu la manipulation et s'attribuent les succès et les erreurs



Les résultats ne changent que peu lorsque les sujets sont informés de la manipulation: ils s'attribuent encore la moitié des « faux succès »!



L'apport de l'électrophysiologie: La négativité liée à l'erreur peut-elle survenir sans conscience?



Immédiatement après une erreur facile à déceler, des ondes cérébrales très caractéristiques sont évoquées:

- *Error-Related Negativity* (ERN)
- suivie d'une positivité centrale (PE)

Ces ondes semblent liées, respectivement, à la détection de l'erreur et à sa correction, ou au minimum son intégration dans le comportement.

La détection et la correction d'erreurs peuvent-elles survenir sans conscience?

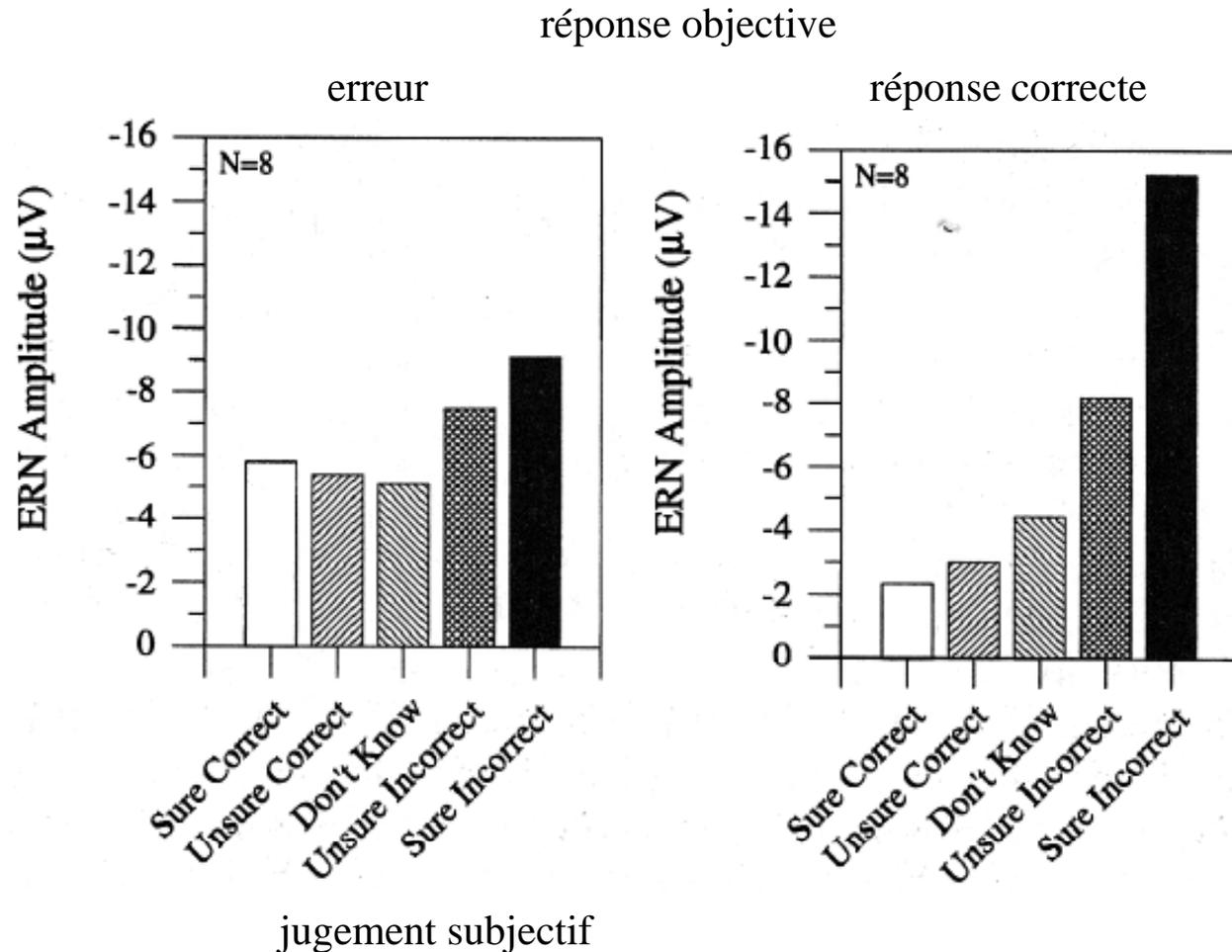
L'ERN reflète un traitement métacognitif: le niveau de confiance des sujets dans leur réponse

Scheffers, M. K., & Coles, M. G. (2000). Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *J Exp Psychol Hum Percept Perform*, 26(1), 141-151.

Dans la tâche des *flankers* d'Eriksen (classifier la lettre centrale dans des stimuli tels que HSHSS):

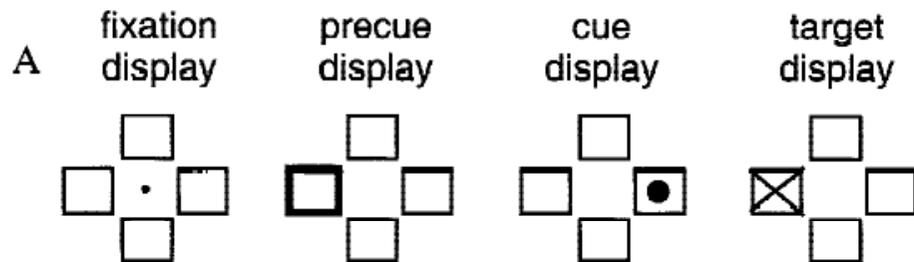
-L'ERN est d'autant plus forte que le sujet est certain d'avoir fait une erreur (et ce, même si la réponse est objectivement correcte!)

- l'ERN reflète un mélange de propriétés objectives et subjectives.



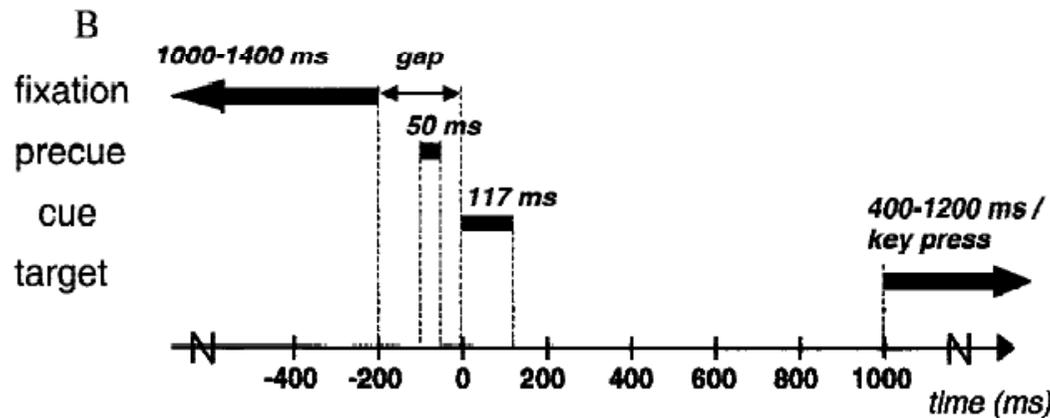
Erreurs non-conscientes dans le paradigme d'anti-saccade

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, 38(5), 752-760.



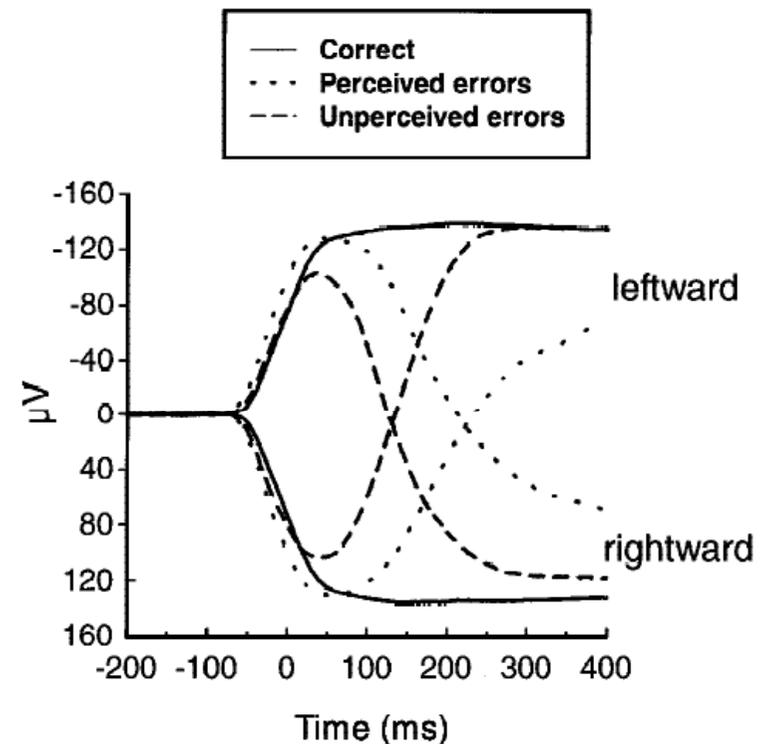
Les sujets doivent déplacer les yeux dans le sens opposé à l'indice (disque noir).

Ensuite, ils indiquent s'ils ont fait une erreur (saccade en direction de l'indice) en appuyant sur la barre d'espace.



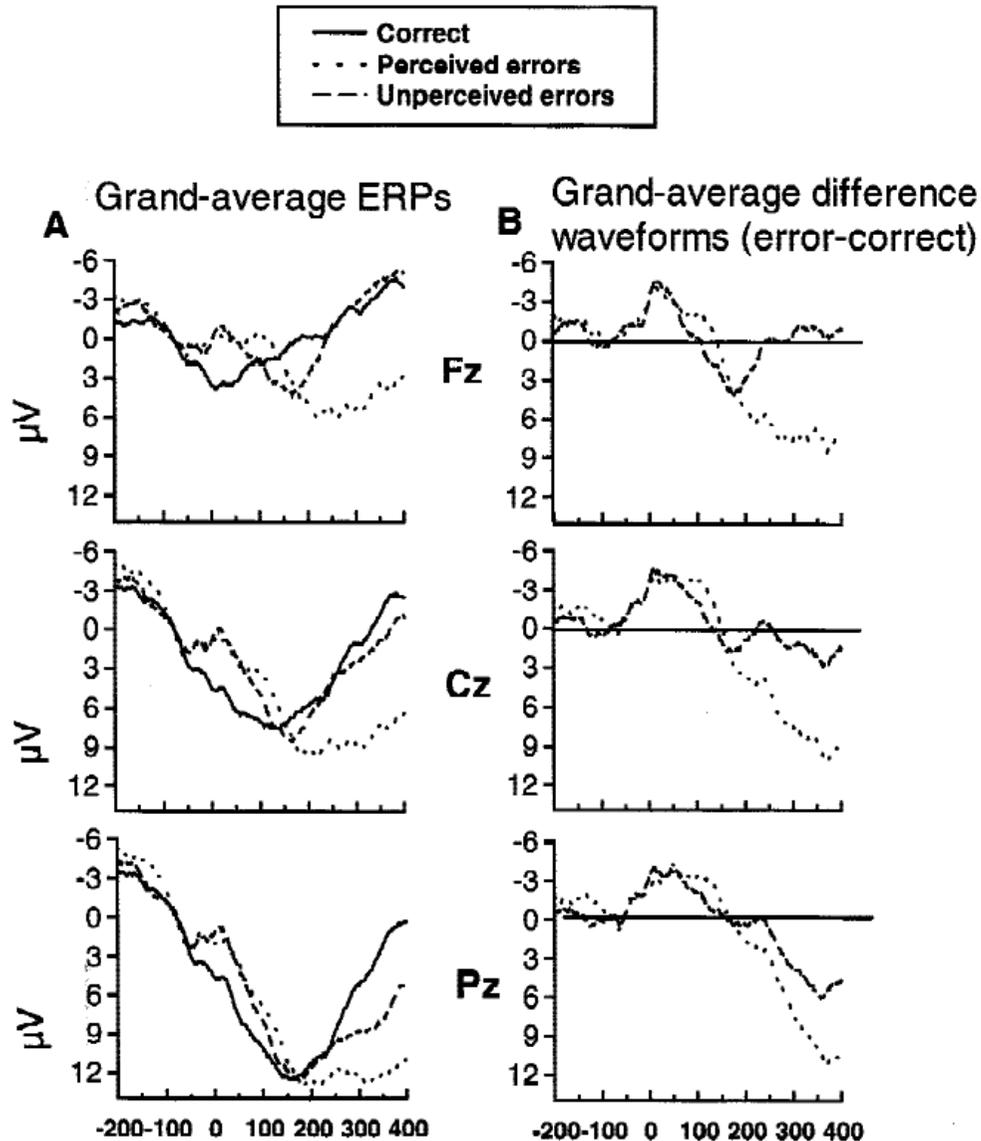
Cette procédure permet de séparer:

- Les essais corrects
- Les erreurs perçues
- Les erreurs non-perçues



Erreurs non-conscientes dans le paradigme d'anti-saccade

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, 38(5), 752-760.



Les résultats montrent deux phases bien séparées:

-Dans la première phase, on observe une négativité (ERN) identique, que le sujet soit ou non conscient de son erreur

-Dans la seconde phase, on n'observe une positivité (Pe) que si le sujet détecte son erreur.

Conclusions: L'ERN indexerait un processus strictement automatique et non-conscient de détection d'erreur, tandis que les ajustements ultérieurs du comportement nécessiteraient la conscience.

Dans certains paradigmes, l'ERN dépend de la conscience de l'erreur

Woodman, G. F. (2010). Masked targets trigger event-related potentials indexing shifts of attention but not error detection. *Psychophysiology, in press*.

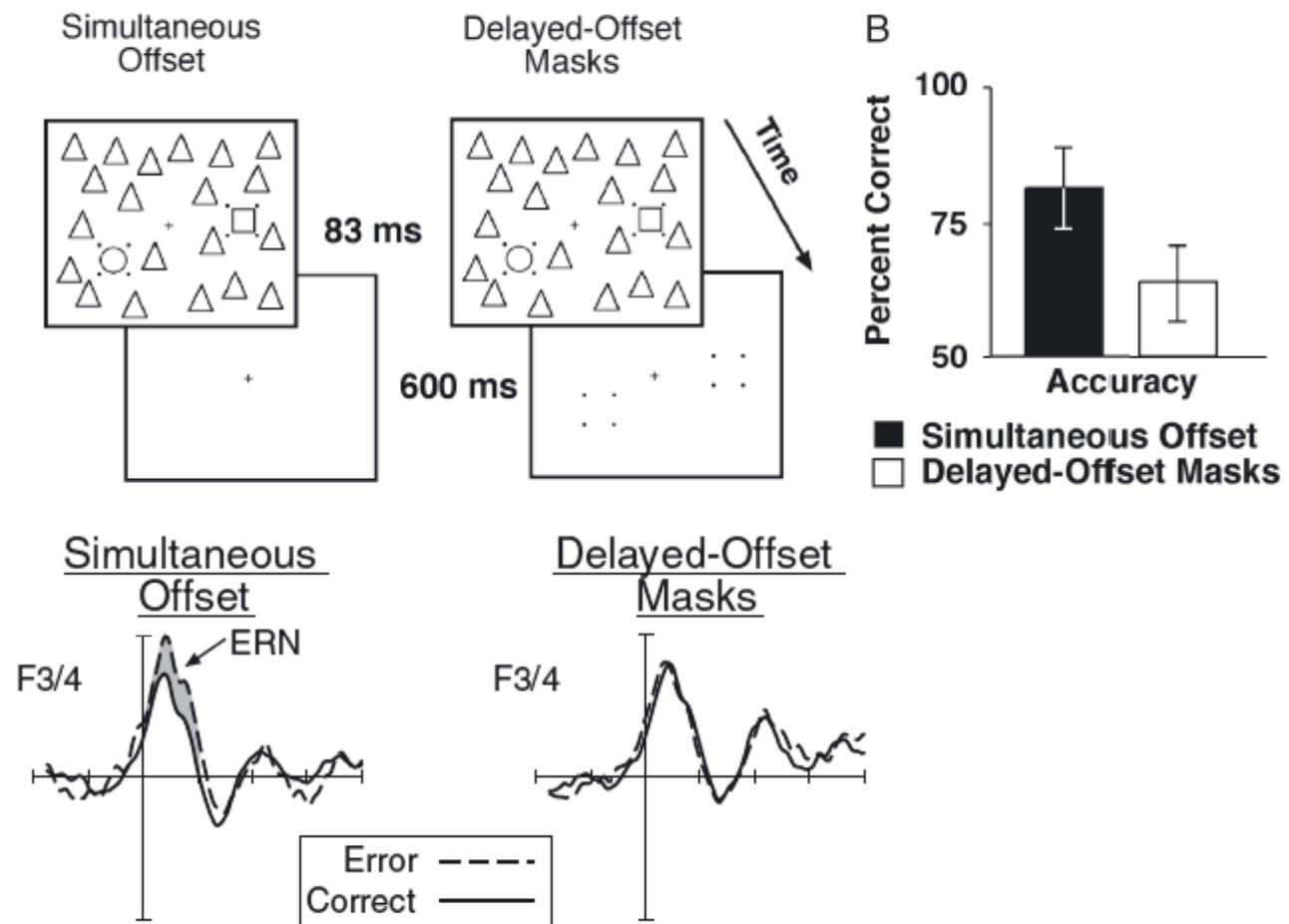
Nieuwenhuis et al. et Logan et al manipulent la conscience de l'action.

Si l'on manipule la conscience du stimulus, on obtient un résultat différent: pas d'ERN non-consciente!

Tâche = recherche d'une cible (carré, losange ou rond, variable selon les blocs)

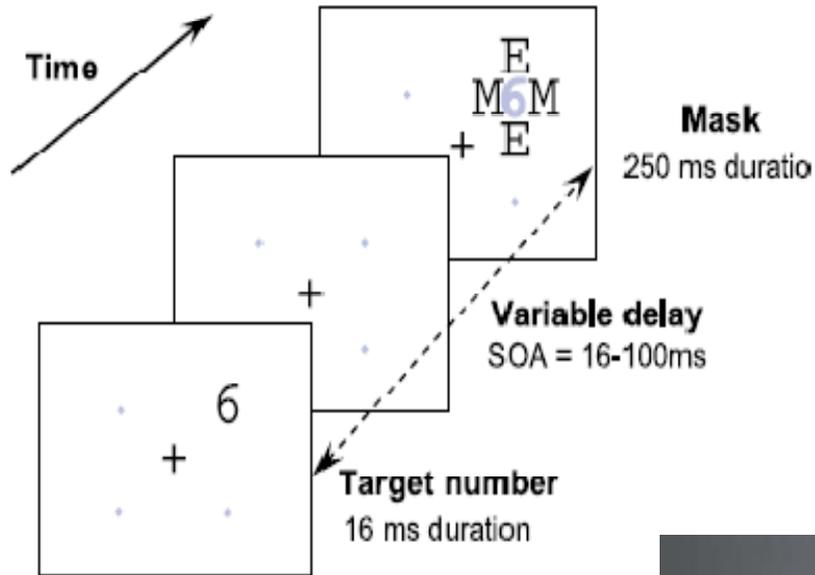
Masquage par substitution: 4 points qui s'interrompent ou continuent après disparition des cibles.

L'ERN n'est pas observée quand les stimuli sont fortement masqués.



Métacognition de l'erreur au cours du masquage

Lucie Charles, thèse de doctorat



3 tâches à chaque essai:

- Comparaison du chiffre avec 5 (avec une pression temporelle forte)
- Visibilité (avez-vous vu le chiffre?)
- Métacognition (avez-vous fait une erreur?)

MEG



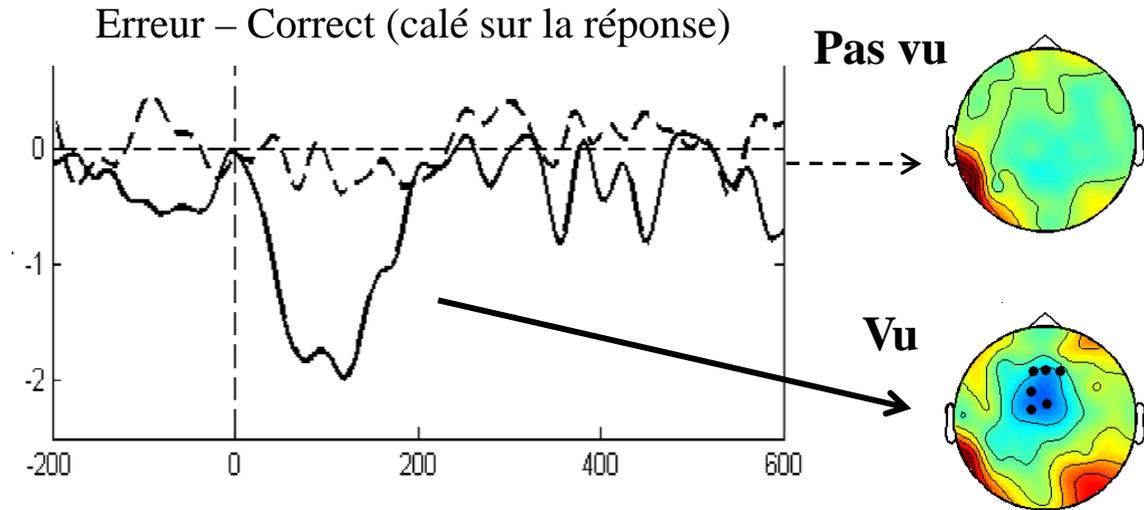
EEG



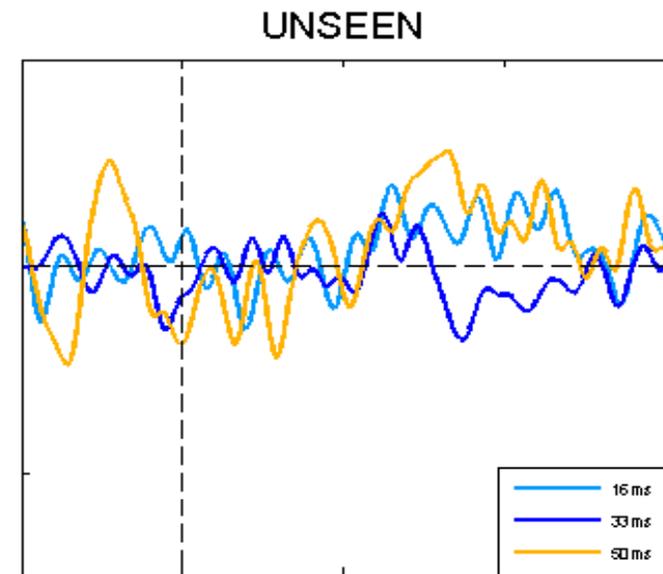
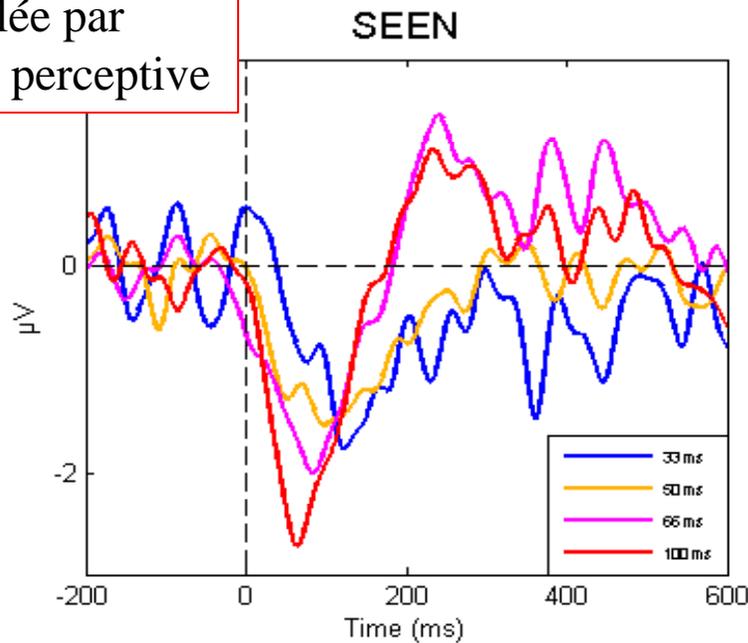
Métacognition de l'erreur au cours du masquage

Lucie Charles, thèse de doctorat

L'ERN n'est présente que lors des essais « vus ».



La pente de l'ERN est modulée par l'évidence perceptive

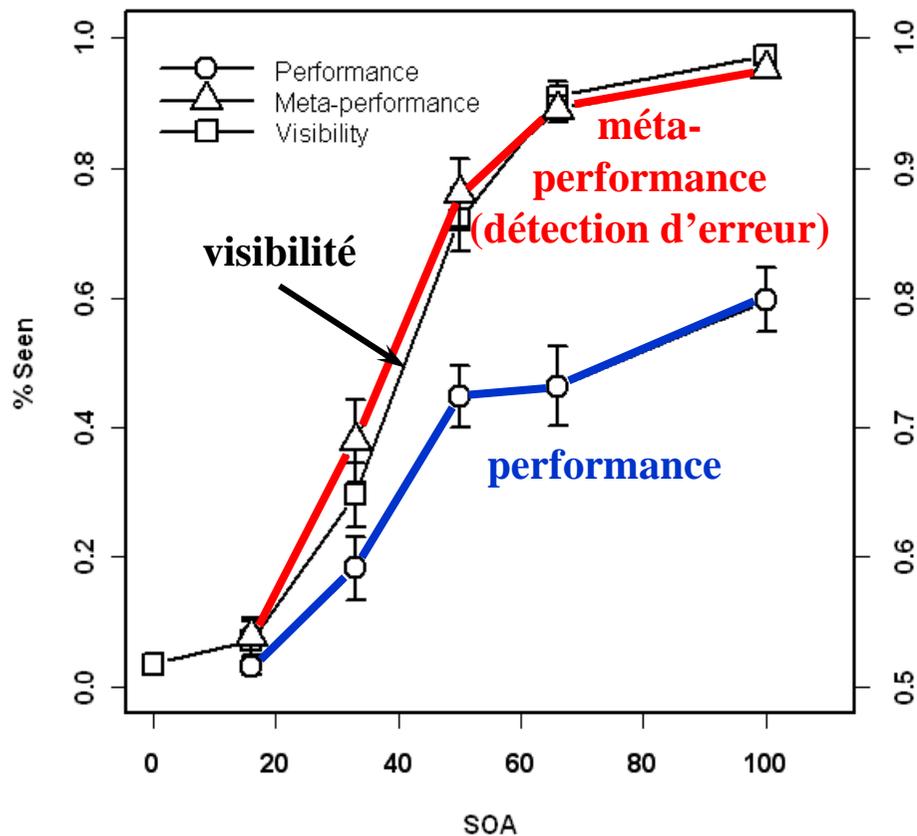


Métacognition de l'erreur au cours du masquage

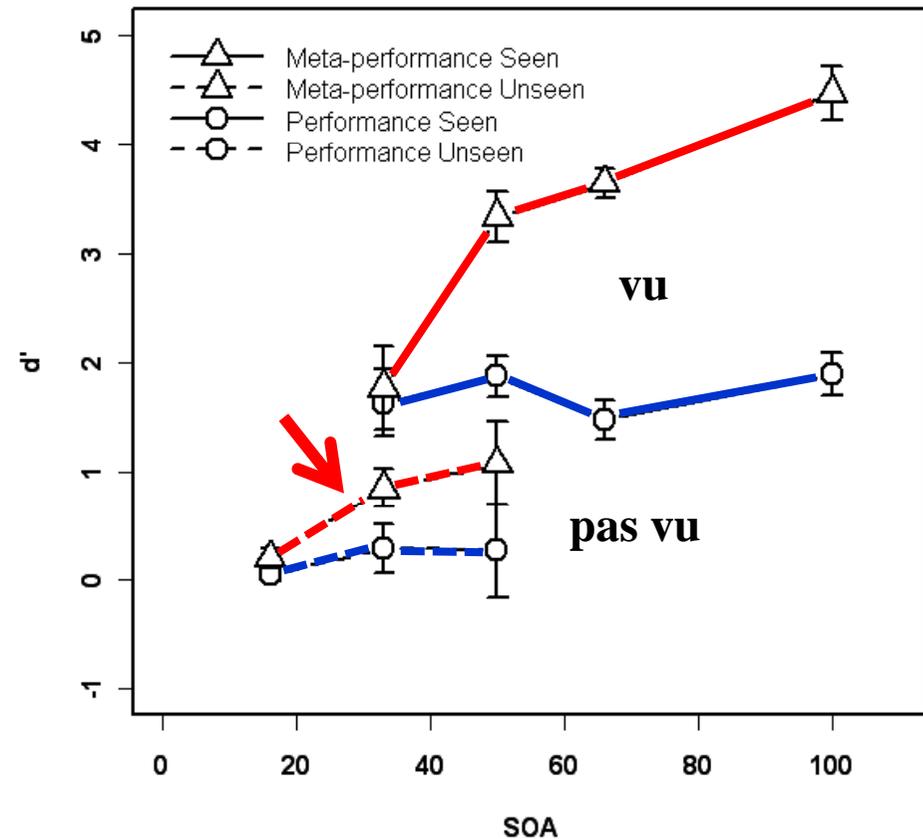
Lucie Charles, thèse de doctorat

- La visibilité et la méta-performance (détection d'erreur) varient de la même manière en fonction du SOA, avec une fonction de seuil
- De plus, à SOA identique, la performance et la méta-performance s'améliorent soudainement dans les essais « vus » par rapport aux essais « pas vus »
- Mais, la méta-performance est meilleure que le hasard même dans les essais « pas vus »!!

Performance, Méta-performance
et Visibilité



d' pour la performance et la méta-performance,
en distinguant essais vus et pas vus



Conclusion: Au moins trois mécanismes, dont deux non-conscients, sont à l'œuvre dans les jugements métacognitifs de confiance et d'erreur

- Une **estimation élémentaire de l'incertitude** accompagne chaque jugement perceptif, même inconscient.

Il se pourrait que chaque aire cérébrale code à la fois

1. le stimulus le plus probable qui explique les entrées sensorielles, ou la réponse la plus probable ou la plus renforcée dans ces circonstances
2. mais également l'incertitude associée à cette estimation
3. et peut-être même toute la distribution de probabilité associée

cf. Galvin mais également les travaux d'Alex Pouget, Sophie Denève, Pascal Mamassian et bien d'autres sur la perception comme inférence Bayésienne

- Il existe également des **systèmes automatisés de détection des erreurs**
 - peut-être même un signal d'*erreur de prédiction* dans chaque aire cérébrale
- En outre, il existe vraisemblablement **un système métacognitif conscient, autonome, et parfois fictif.**
 - Auto-attributions des erreurs dans la tâche de Logan & Crump (2010)
 - variation inter-individuelle indépendante de la performance à choix forcé (Fleming et al. *Science* 2010)
 - système cérébral distinct (cortex préfrontal, aire 10 de Brodmann)