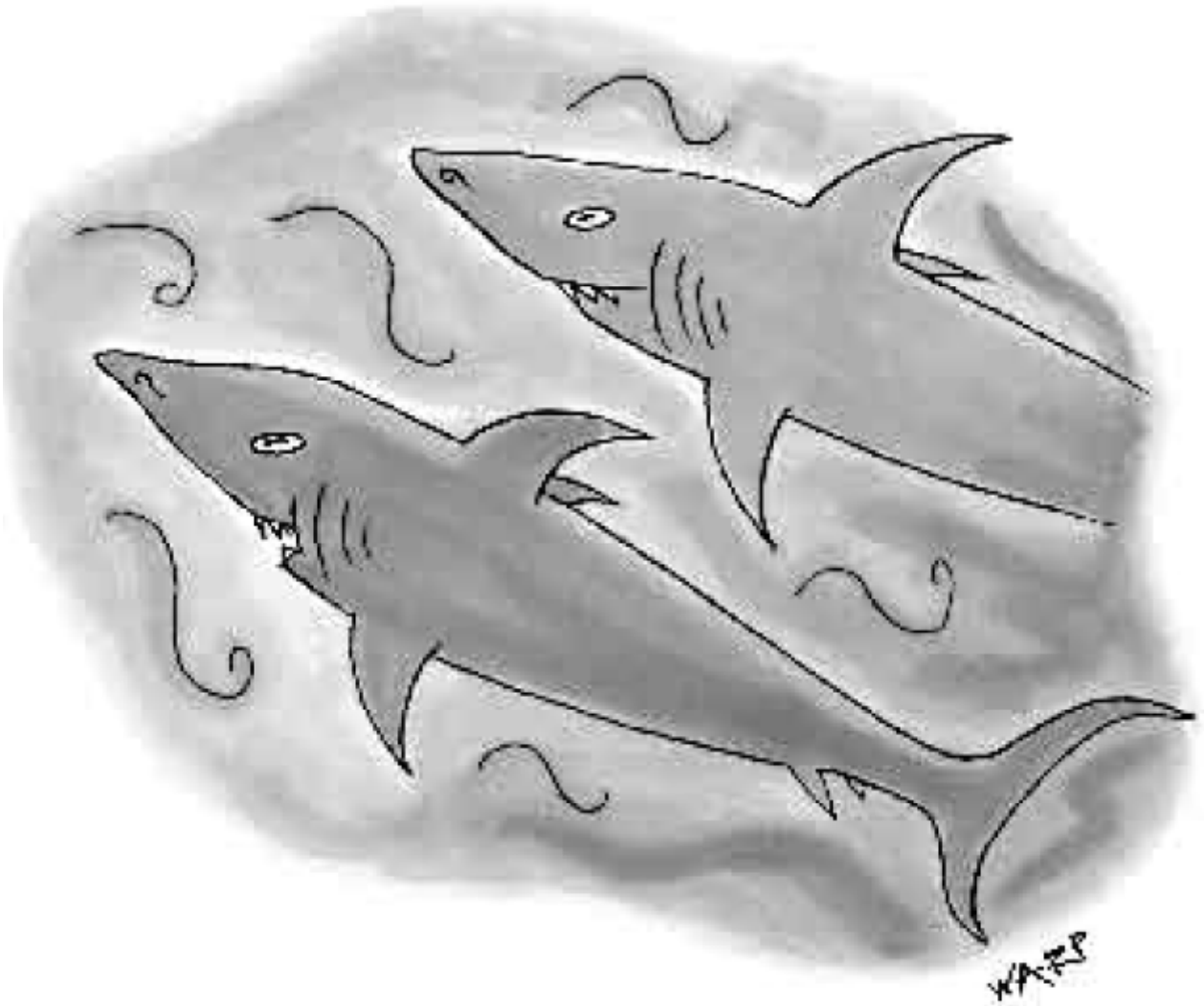
Introspection et métacognition :
Les mécanismes de la connaissance de soi

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

Cours

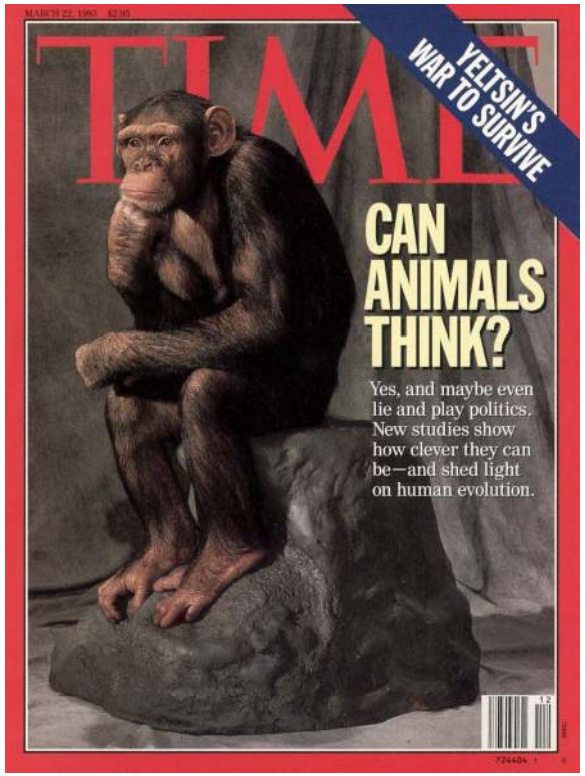
Modèles expérimentaux de l'introspection chez l'animal



Ah, j’adore l’odeur de la crème solaire au petit matin …

Les animaux disposent-ils d'une forme d'introspection?

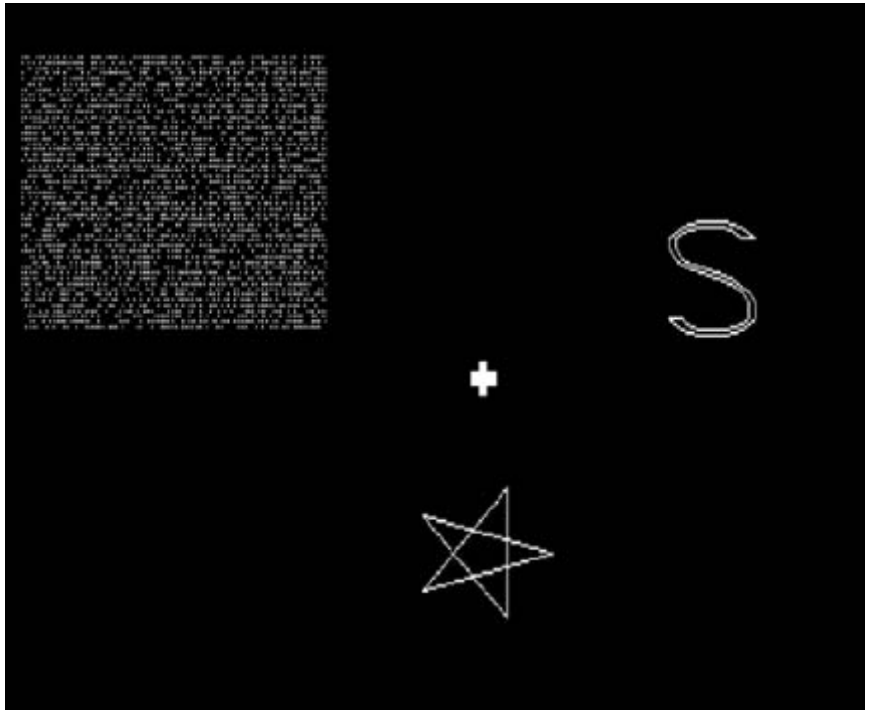
- De nombreuses expériences de métacognition reposent sur un rapport verbal ou numérique (degré de confiance dans sa réponse).
- Les compétences métacognitives ne sont typiquement pas observées avant l'âge de 4 ans, ce qui a conduit certains à postuler que la métacognition requiert le langage (Metcalfe & Shimamura, 1994; Tulving, 1994).
- Toutefois, nous avons vu dans le cours précédent que des méthodes non-verbales permettent d'attribuer une théorie de l'esprit aux enfants de 14 et même de 7 mois.
- Kornell, Son et Terrace (2007) proposent au moins deux manières de tester la métacognition sans langage:
    1. Introspection (*metacognitive monitoring*): l'animal peut-il rapporter son degré de confiance dans ses propres réponses, par exemple en choisissant une réponse risquée ou moins risquée?
    2. Contrôle métacognitif: l'animal peut-il montrer qu'il « sait qu'il ne sait pas », par exemple en allant activement chercher des informations supplémentaires?



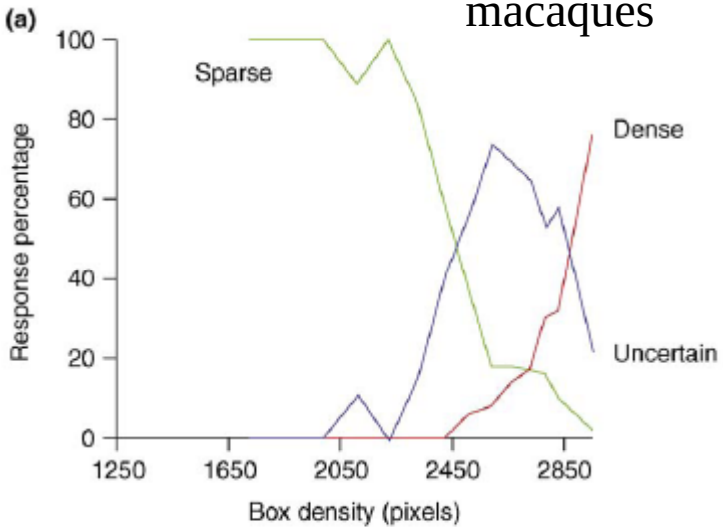
Le paradigme de Smith: la réponse « incertaine »

Smith, J. D., Shields, W. E., Schull, J., & Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition, 62(1), 75-97.*

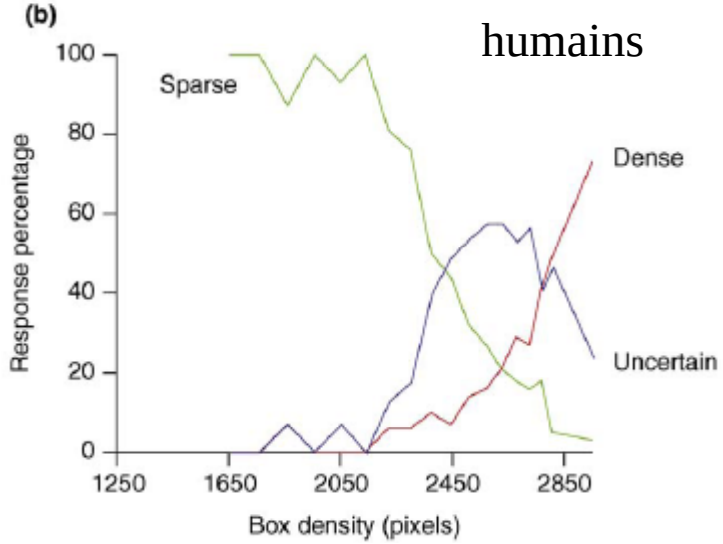
- Jugement de densité d'un écran
- La touche « étoile » permet d'échapper à la tâche primaire et de recevoir une récompense modeste mais fixe.
- L'animal peut ainsi manifester son « incertitude » ou son « manque de confiance en soi »



macaques



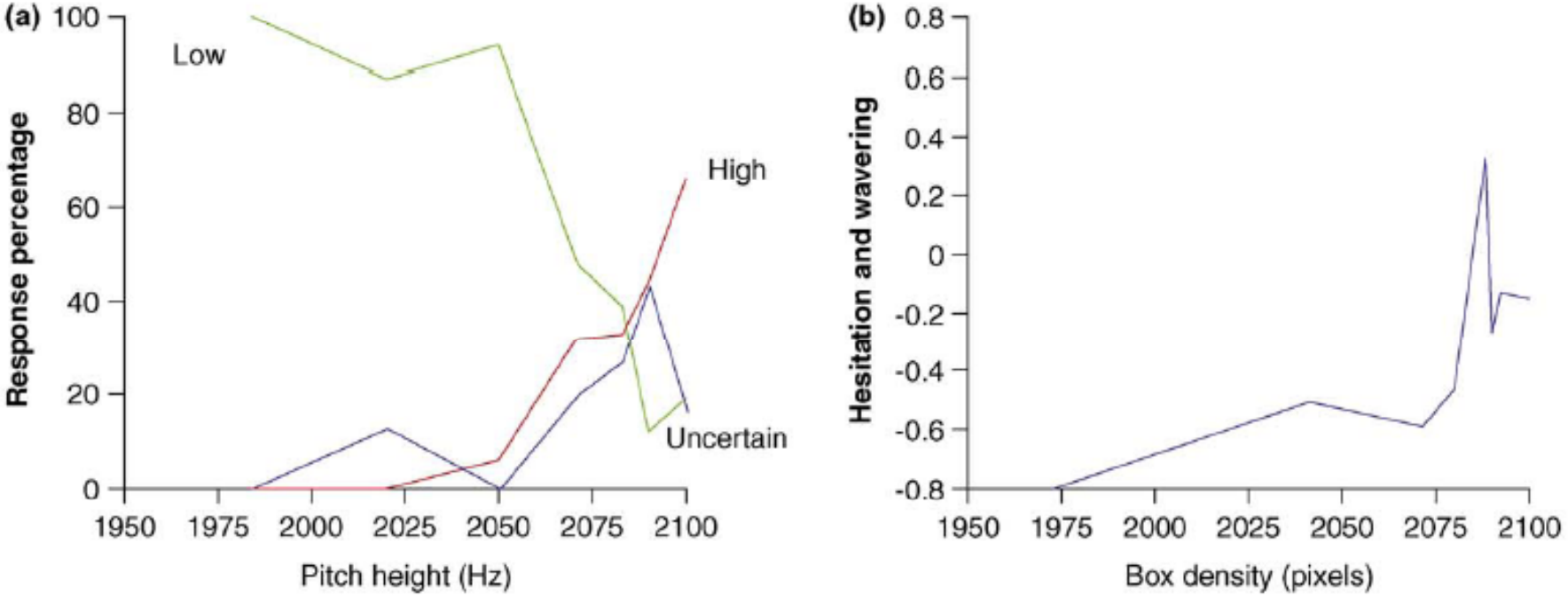
humains



Le paradigme de Smith chez le dauphin

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., & Erb, L. (1995). The uncertain response in the bottlenosed dolphin (Tursiops truncatus). *J Exp Psychol Gen, 124(4), 391-408.*

- Jugement de la hauteur d'un son: s'agit-il de 2100 Hz ou d'une fréquence plus basse?
- L'animal peut refuser l'essai et reçoit alors un essai facile.
- La discrimination est excellente, mais l'animal refuse spécifiquement les essais difficiles.
- L'animal présente également des réponses d'hésitation (nage lentement, secoue la tête!), précisément dans la même zone de fréquence.



La classification critique de Terrace et Son (2009)

Experiment	Species	Task	Judgment
Category 1			
Smith *et al.* [19]	Monkey	Psychophysical	Concurrent[a]
Shields *et al.* [20]	Monkey	Psychophysical	Concurrent
Smith *et al.* [21]	Monkey/humans	SPR[b]	Concurrent
Inman and Shettleworth [22]	Pigeon	DMTS[c]	Concurrent
Call and Carpenter [23]	Ape/children	Hidden objects	Concurrent
Shields *et al.* [24]	Monkey	Psychophysical	Concurrent
Beran *et al.* [25]	Monkey	Psychophysical	Concurrent
Suda-King [26]	Orangutan	Spatial Memory	Concurrent
Washburn *et al.* [27]	Monkey	Psychophysical/MTS[d]	Concurrent
Sutton and Shettleworth [28]	Pigeon	MTS/DMTS	Concurrent
Basile *et al.* [29]	Monkey	Hidden objects	Concurrent
Kepecs *et al.* [30•]	Rat	Psychophysical	Concurrent
Foote and Crystal [31]	Rat	Psychophysical	Concurrent
Category 2[e]			
Smith *et al.* [33]	Monkey	Psychophysical	Prospective
Category 3[f]			
Hampton [34••]	Monkey	DMTS	Prospective
Category 4			
Kornell *et al.* [35••]	Monkey	SPR	Retrospective

[a] Concurrent metacognition is based on judgments made in the presence of the discriminative stimuli that are presented during a particular trial. They should be distinguished from prospective metacognition, which is based on confidence to perform accurately on an upcoming test and, retrospective metacognition, which is based on confidence in the accuracy of their response on the present trial.

[b] SPR = serial probe recognition (see p. 7).

[c] DMTS = delayed-matching-to-sample (see p. 6).

[d] MTS = matching to sample.

[e] Category 2 includes experiments that claim that there was no reinforcement for escape responses, but that claim can be questioned on two grounds. These studies used subjects that were trained previously to use escape responses that were reinforced and it has been shown that metacognitive skills transfer readily to new tasks [35••].

[f] Although food reward was provided in experiments in Categories 3 and 4, none of the discriminative stimuli were present during the delivery of the reward. In the absence of those stimuli, subjects could opt for large rewards if they were confident about their ability to solve a problem that they could take during an upcoming test (prospective metacognition) or about their performance on the trial that had just ended (retrospective metacognition).

Les singes savent quand ils se souviennent

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proc Natl Acad Sci U S A, 98(9), 5359-5362.*

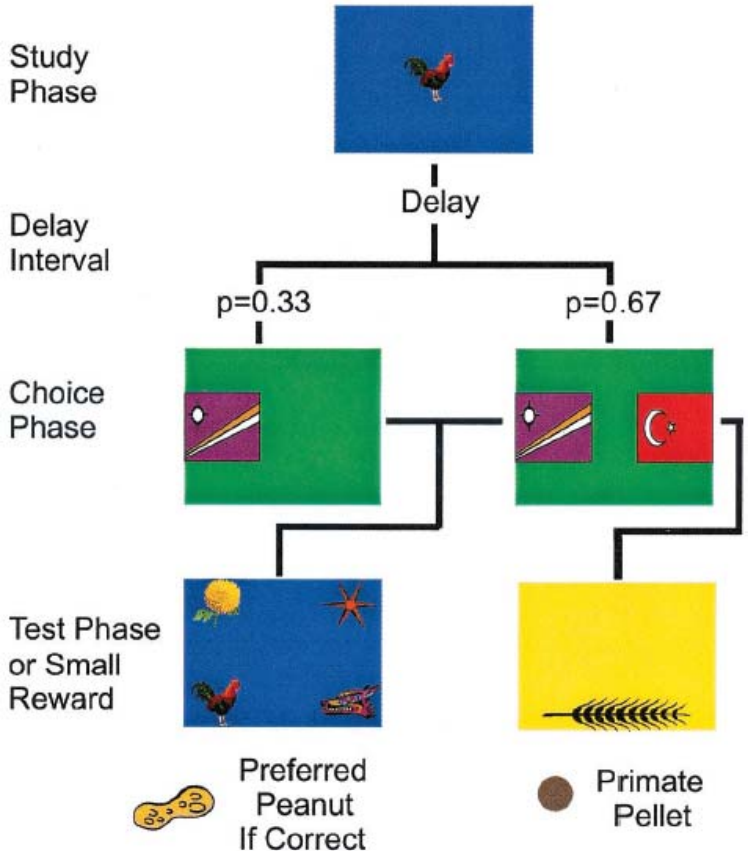
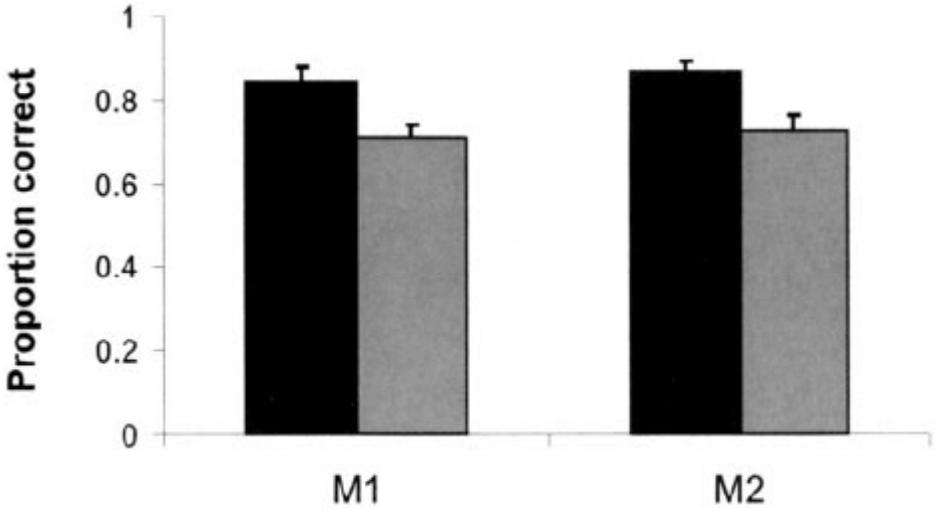


Fig. 1. Method for assessing whether monkeys know when they remember. Each colored panel represents what monkeys saw on a touch-sensitive computer monitor at a given stage in a trial. At the start of each trial, monkeys studied a randomly selected image. A delay period followed over which monkeys often forgot the studied image. In two-thirds of trials, animals chose between taking a memory test (*Right*, left-hand stimulus) and declining the test (*Right*, right-hand stimulus). In one-third of trials, monkeys were forced to take the test (*Left*). Better accuracy on chosen than on forced tests indicates that monkeys know when they remember and decline tests when they have forgotten, if given the option.

Métacognition *prospective* (avant d'avoir terminé la tâche primaire):
L'animal doit mémoriser une image.
Dans certains essais, un test de mémoire lui est imposé (à gauche).
Dans d'autres, il peut choisir de le refuser, auquel cas il reçoit une petite récompense fixe.

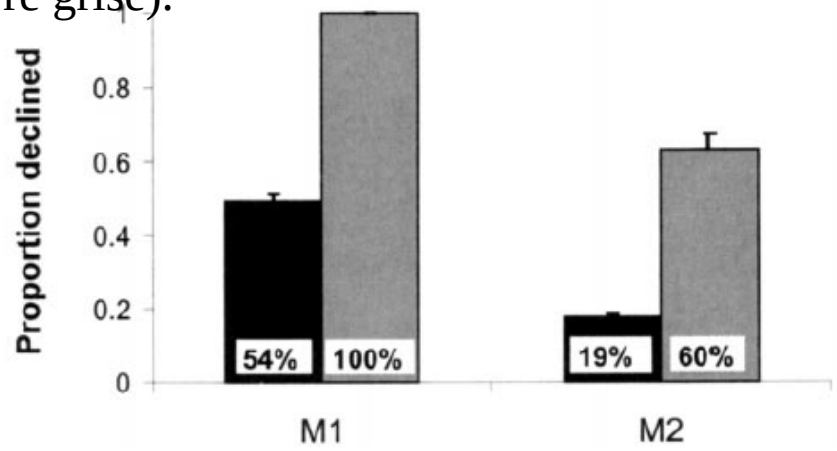


Chez deux singes, la performance de mémoire est meilleure quand le singe choisir de passer le test de mémoire (barres noires) que quand le test lui est imposé (barres grises)

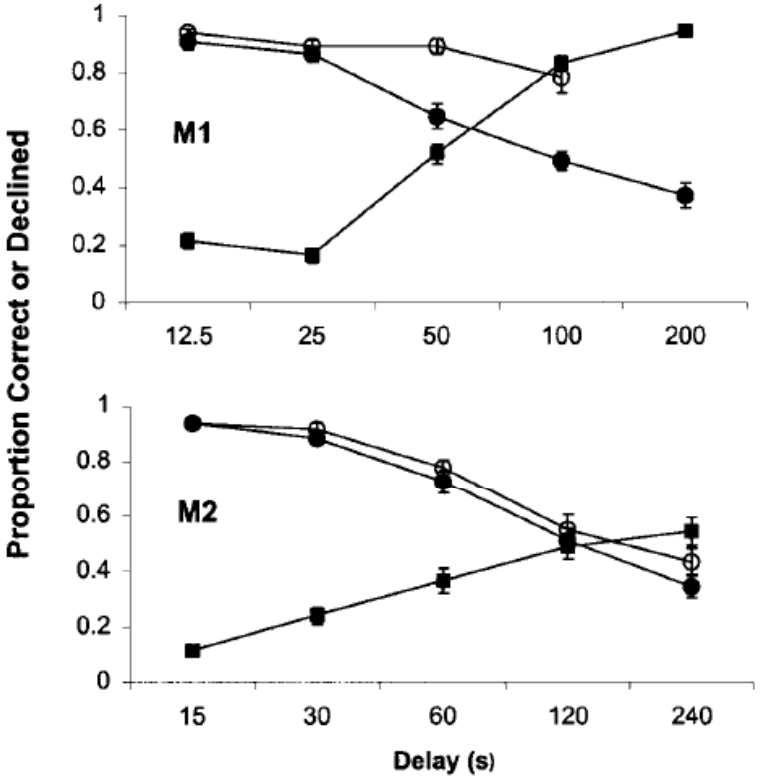
Les singes savent quand ils se souviennent

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proc Natl Acad Sci U S A, 98(9), 5359-5362.*

Contrôle expérimental (expérience 2):
Dans une fraction des essais, aucune image n'est présentée durant la phase de mémorisation. Si les singes comprennent la tâche, il doivent toujours refuser le test de mémoire. Si, par contre leur performance se fonde sur d'autres indices comportementaux (mouvement, distraction auditive…) alors ils devraient se comporter de la même manière dans les essais avec et sans image.
Résultat: les singes refusent beaucoup plus souvent lors des essais sans image à mémoriser (barre grise).



Expérience 3:
Variation du délai de mémoire:
Les animaux refusent le test de plus en plus souvent à mesure que le délai augmente. Leur performance est meilleure quand ils ont le choix de refuser.



Une authentique démonstration de métacognition chez l'animal

Kornell, Son & Terrace, *Psychological Science* 2007

- Métacognition *rétrospective*: après avoir agi, mais avant d'avoir reçu une récompense, l'animal doit juger de ses propres performances.
- Deux critères nouveaux pour une authentique démonstration de métacognition chez l'animal
- Critère 1: l'animal ne doit pas seulement apprendre à utiliser la réponse d'évitement à bon escient, mais il doit **généraliser** cette réponse à des tâches nouvelles
- La généralisation montrerait que l'animal utilise effectivement la réponse à renforcement fixe pour manifester son degré de confiance dans sa réponse – et qu'il ne s'agit donc pas d'une réponse comme les autres, évoquée par un type restreint de stimuli.
- Dans l'expérience 1 de Kornell et al. (2007), deux singes macaques

  1. apprennent à utiliser à bon escient la réponse à renforcement fixe dans une tâche de jugement de taille

  2. généralisent immédiatement cette réponse à une nouvelle tâche de mémoire.

Une authentique démonstration de métacognition chez l'animal

Kornell, Son & Terrace, *Psychological Science* 2007

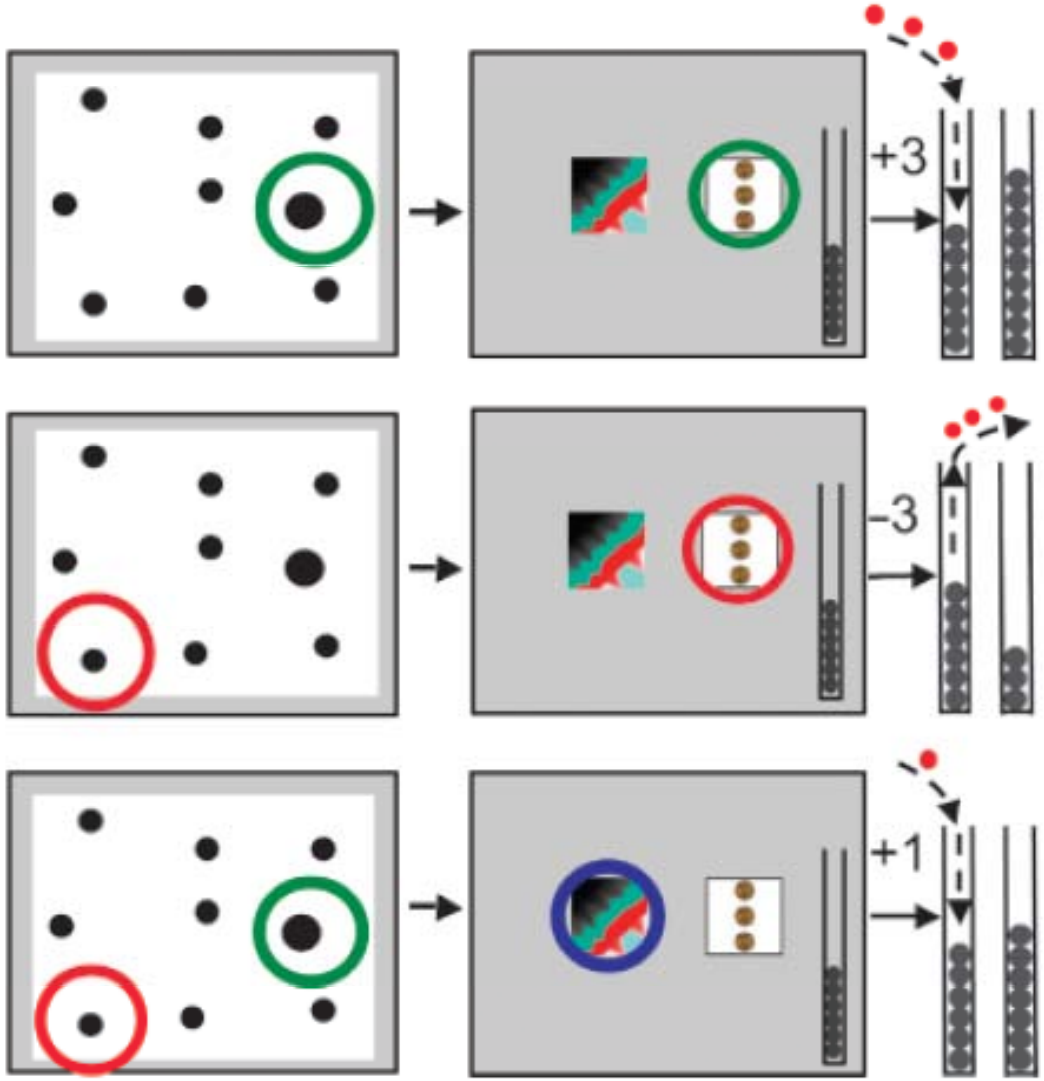
Tâche primaire = Jugement de taille: choisir le plus grand disque.

Tâche secondaire = prise de risque

L'animal apprend que

1. s'il choisit bien, et qu'il clique sur l'icône de droite, il obtient 3 croquettes de nourriture.
2. s'il choisit mal, et qu'il clique sur l'icône de droite, on lui retire 3 croquettes.
3. s'il choisit l'icône de gauche, il reçoit, quoi qu'il ait choisi, une seule croquette.

Noter que la réponse métacognitive est faite une fois que les stimuli de la tâche primaire ont disparu.



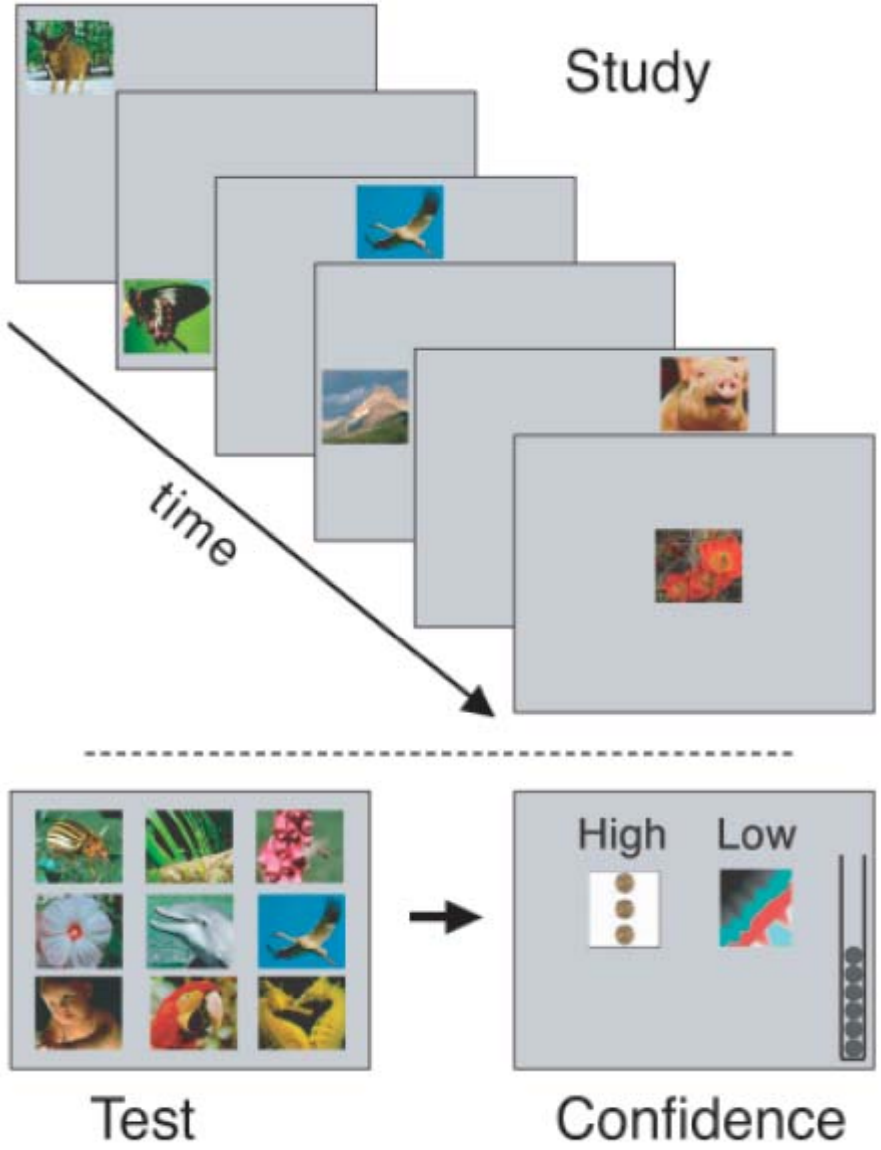
Une authentique démonstration de métacognition chez l'animal

Kornell, Son & Terrace, *Psychological Science* 2007

Tâche primaire numéro 2 = Mémoire à court terme: Visualiser une série de six images, puis choisir, parmi 9 images, laquelle était présente dans la liste précédente.

L'animal apprend d'abord à faire la tâche seule.

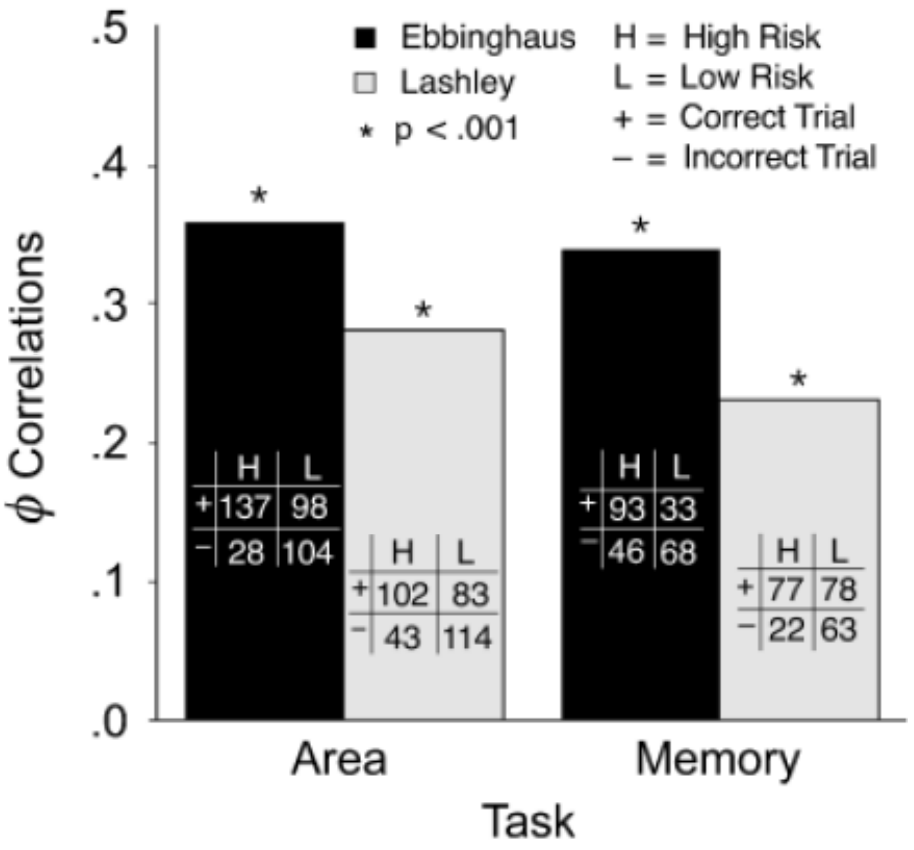
Puis on introduit le jugement supplémentaire de prise de risque (plus d'une vingtaine de sccondes après les stimuli).



Une authentique démonstration de métacognition chez l'animal

Kornell, Son & Terrace, *Psychological Science* 2007

Les résultats sont quantifiés par la corrélation (phi) entre la réussite à la tâche primaire, et la prise de risque.
On observe
- une corrélation significative, chez les deux animaux testés:
-une généralisation presque parfaite à l'autre tâche primaire (mémoire).
Les animaux utilisent-ils un indice élementaire? Non:
- la corrélation reste présente quelle que soit la position sérielle de la cible dans la liste; le jugement des animaux n'est donc pas fondé sur cet indice
- la performance métacognitive est négativement corrélée avec le temps de réaction (les erreurs sont plus lentes), mais elle reste significative même quand on retranche l'effet du temps de réaction par régression.



Une authentique démonstration de métacognition chez l'animal

Kornell, Son & Terrace, *Psychological Science* 2007

- Critère 2 = capacité de **contrôle métacognitif:**

L'animal doit être capable d'utiliser son degré de confiance pour rechercher activement des informations nouvelles.

- Dans des expériences antérieures (Call & Carpenter, 2001; Hampton, Zivin, & Murray, 2004), si les animaux ont vu cacher de la nourriture, ils vont directement à l'endroit approprié, alors que si ne l'ont pas vu, ils la recherchent activement.
  - Pour Terrace, cela ne signifie pas qu'ils « savent qu'ils ne savent pas ». Récupération et recherche de nourriture sont des comportements spontanés, déclenchés directement par le fait de savoir où est la nourriture (ou ne pas le savoir), ce qui n'est pas une métacognition (« savoir que l'on sait »).
- Dans l'expérience 2
  - les animaux apprennent une séquence nouvelle d'images par essai et erreurs, avec une touche supplémentaire qui leur donne un indice.
  - Les résultats montrent qu'ils utilisent cette touche à bon escient.

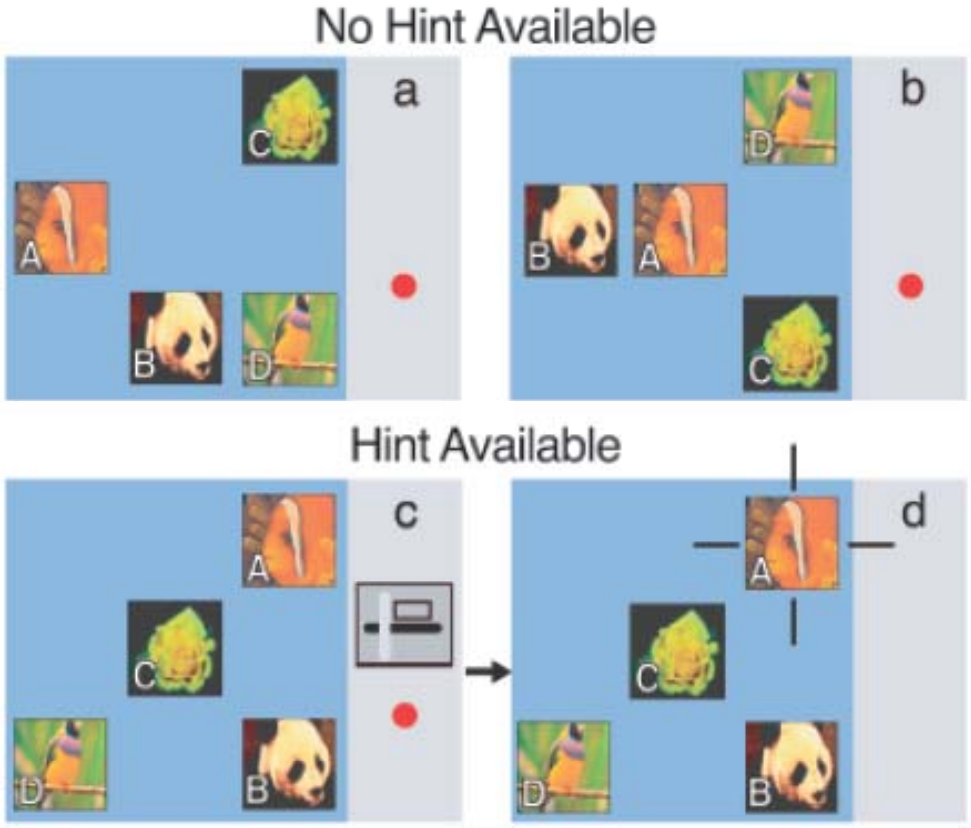
Une authentique démonstration de métacognition chez l'animal

Kornell, Son & Terrace, *Psychological Science* 2007

Tâche =
découvrir l'ordre dans lequel une série d'images doit être ordonnée (en appuyant sur un écran tactile).

Dans la moitié des essais, une touche supplémentaire « indice » est disponible.
Un appui sur cette touche indique à l'animal quel est le prochain élément de la séquence.

Cela conduit également à une dévaluation de la récompense:
-le point rouge indique que l'animal recevra un M&M en cas de réussite
- l'appui sur la touche « indice » élimine ce point rouge: la récompense ne sera qu'une croquette à la banane.



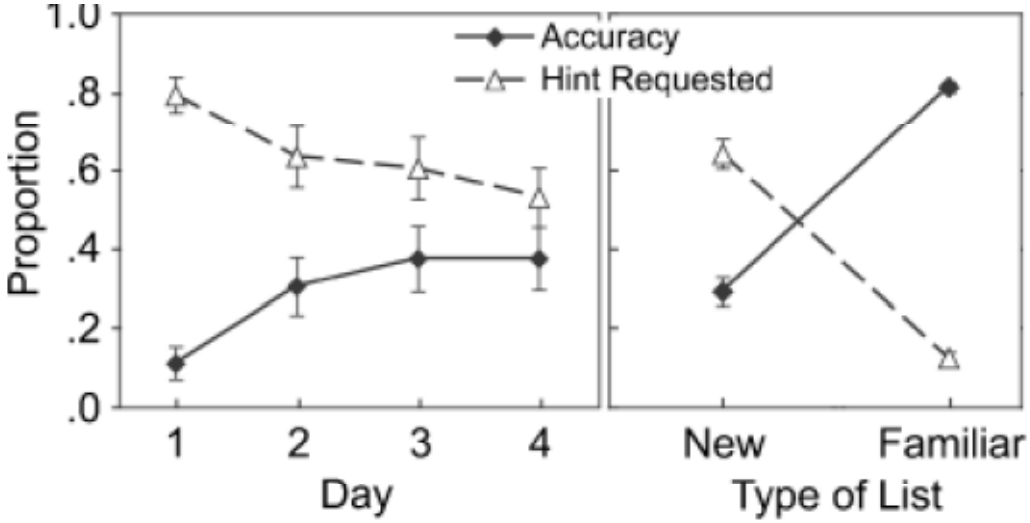
Une authentique démonstration de métacognition chez l'animal

Kornell, Son & Terrace, *Psychological Science* 2007

Résultats:

-A mesure que chaque liste est apprise, la proportion d'appuis sur la touche « indice » diminue proportionnellement.

- Cette corrélation négative existe au sein de chaque liste, et aussi bien dans les listes familières que dans les listes nouvelles – mais l'introduction d'une liste nouvelle augmente considérablement le nombre d'indices demandés.



+ C'est la toute première expérience qui étudie la recherche active d'informations chez l'animal

- Cette expérience est un peu moins convaincante que la précédente :

- L'indice est-il vraiment « compris » par l'animal comme signifiant qu'un complément d'information est disponible? Une expérience de généralisation aurait permis de trancher.

- Est-il vraiment impossible d'interpréter cette expérience en termes de maximisation de la récompense? Lorsque la liste est nouvelle, l'animal obtient une récompense plus rapidement en utilisant les indices; lorsque la liste est ancienne, il obtient une récompense plus élevée s'il ne l'utilise pas. Il se pourrait que l'animal se serve du temps écoulé comme une indication de la « nouveauté » de la liste, et que la variable « je sais que je connais cette liste » ne soit pas utilisée dans la décision. Corrélation n'est pas causation!

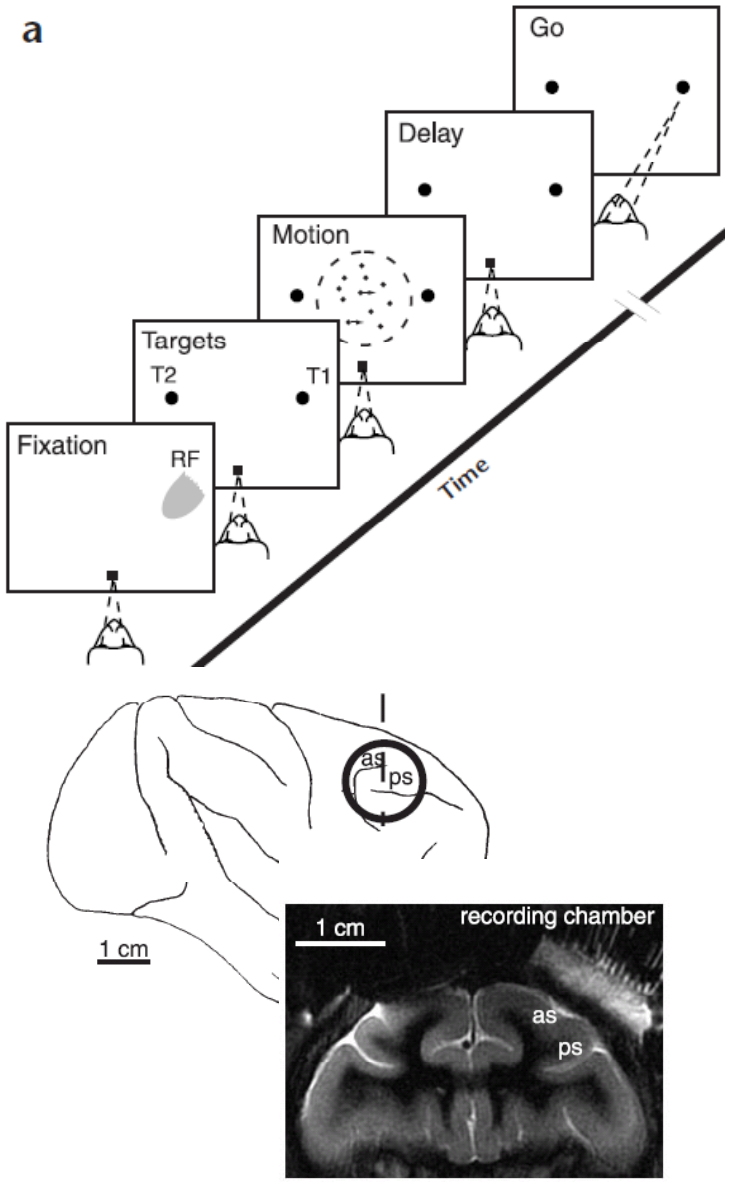
Vers une neurophysiologie de l'introspection

Les protocoles métacognitifs qui évaluent le degré de confiance en soi de l'animal sont devenus suffisamment simples pour conduire à une expérimentation neurophysiologique.

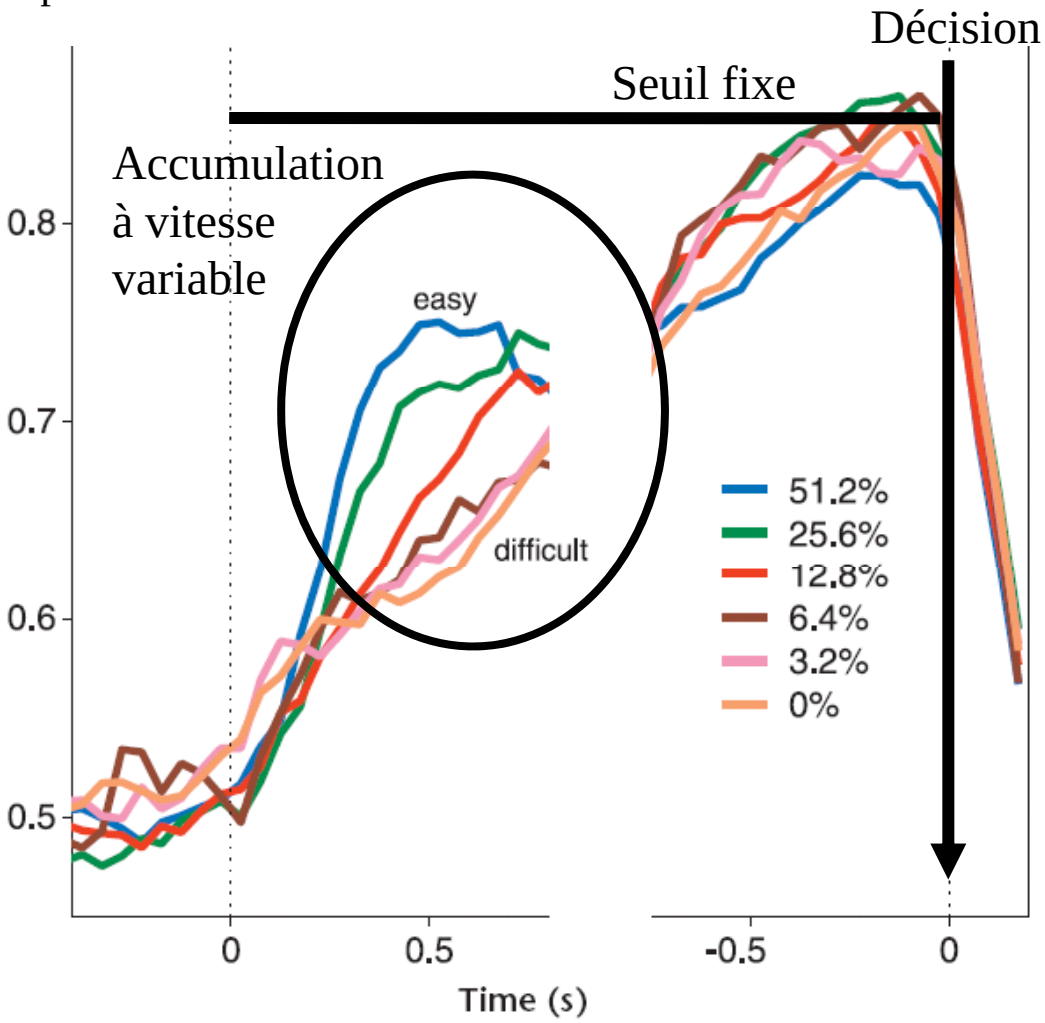
Deux articles très récents sur ce sujet:

- Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324(5928), 759-764.*
- Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455(7210), 227-231.*

Les neurones préfrontaux et pariétaux compilent des statistiques pertinentes pour la prise de décision

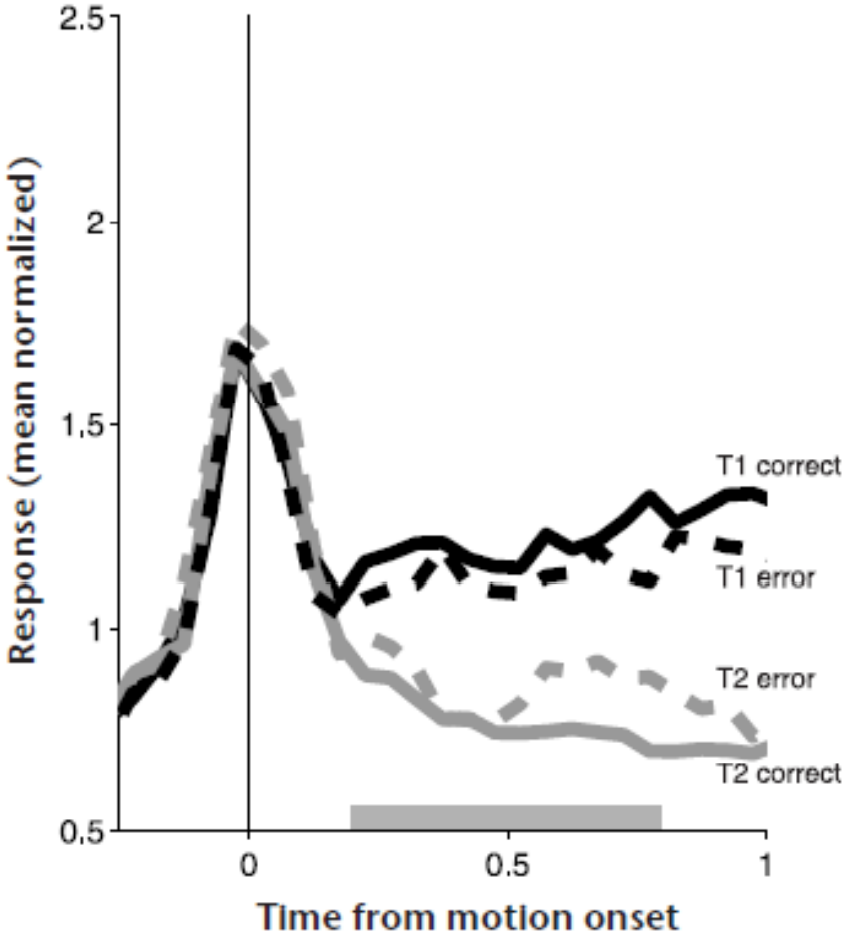


Le singe doit décider de la direction principale du mouvement. La cohérence du mouvement module l'augmentation des réponses neuronales au cours de la décision.



Kim & Shadlen, *Nature Neuroscience* 1999

Les signaux neuronaux associés à la prise de décision contiennent également des informations sur la confiance que l'on peut lui accorder.



En particulier, la réponse des neurones au cours de la décision distingue les essais corrects et les essais erronés.

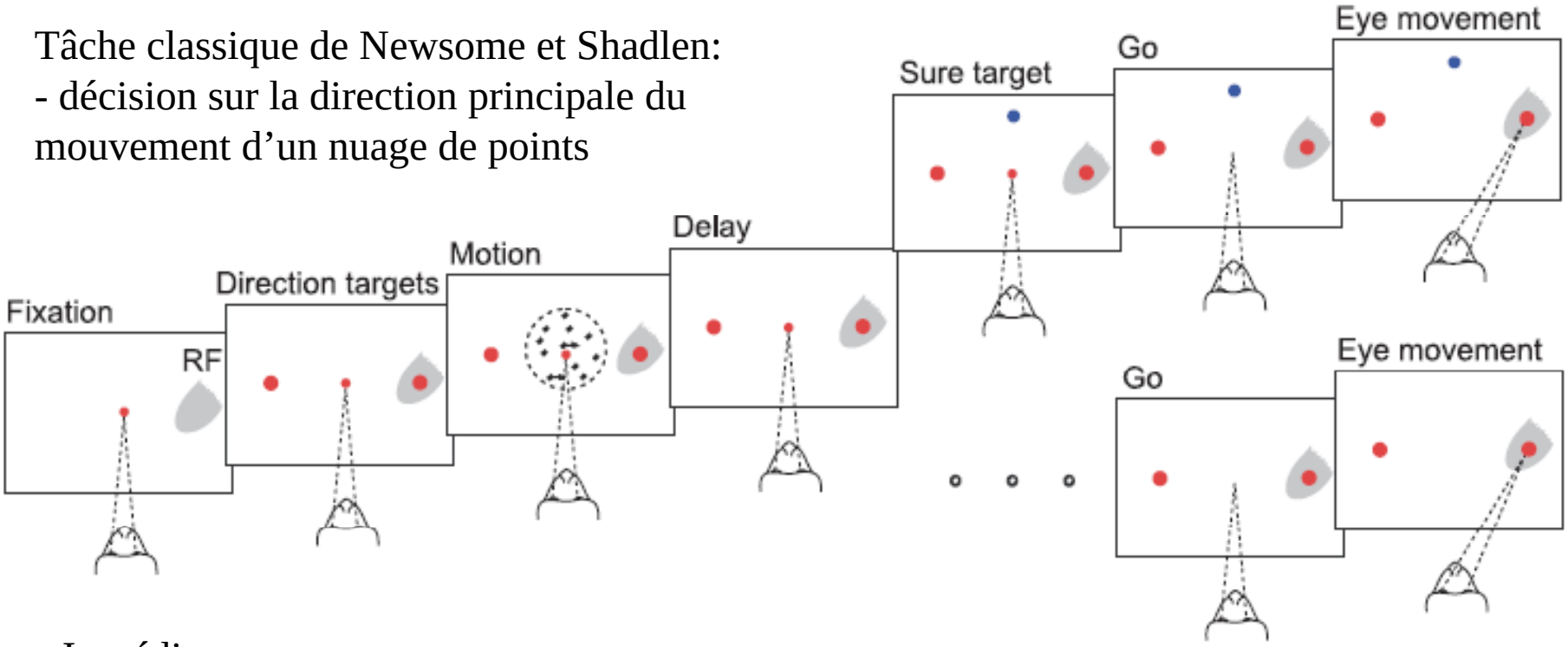
La différence d'activité des neurones codant pour les deux décisions possibles pourrait donner une indication sur la confiance que l'on peut accorder à la décision (mécanisme proposé par Vickers, 1979)

Kim & Shadlen, *Nature Neuroscience* 1999

Représentation de la confiance par les neurones du cortex pariétal

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324(5928), 759-764.*

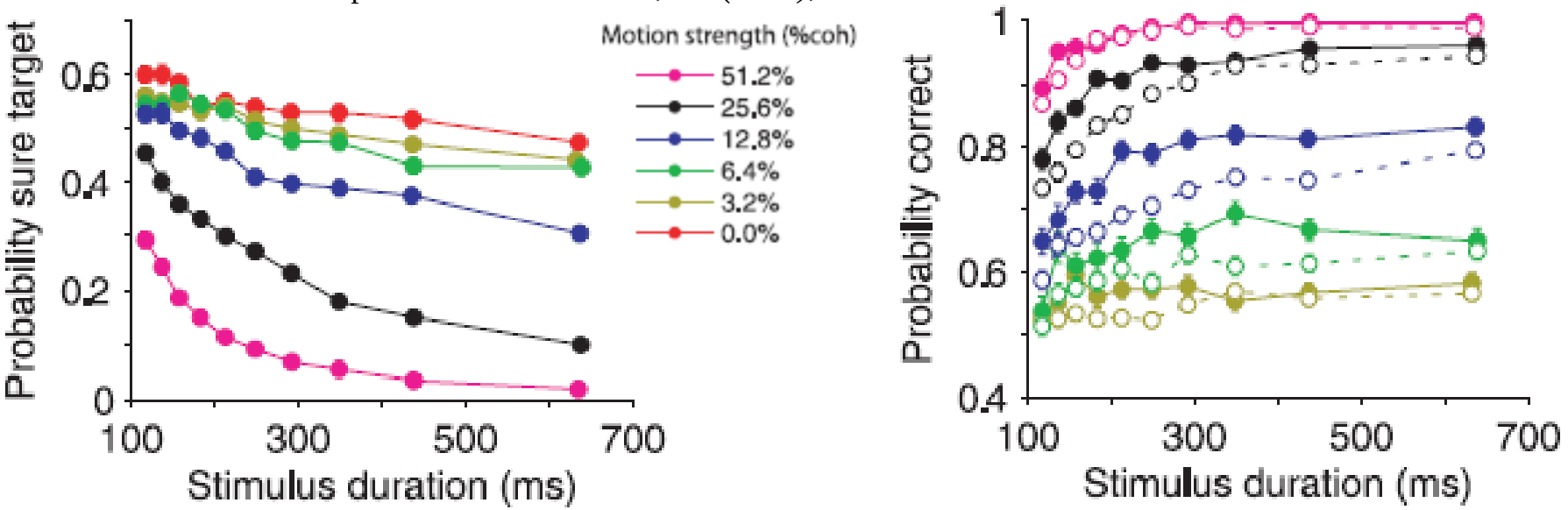
Tâche classique de Newsome et Shadlen:
- décision sur la direction principale du mouvement d'un nuage de points



Ingrédient nouveau:
- *avant* la décision, une troisième cible apparaît à certains essais (*sure target*)
- le choix de cette cible conduit à une récompense fixe mais moindre.
- ce choix peut être interprété comme un refus de répondre à la tâche principale, ce qui pourrait indiquer que l'animal est « incertain », n'a pas « confiance en lui ».

Le comportement reflète une représentation appropriée de la confiance de l'animal en son propre jugement

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324(5928), 759-764.*



Le singe choisit de refuser de répondre d'autant plus souvent que la probabilité d'une réponse correcte à la tâche primaire est petite.

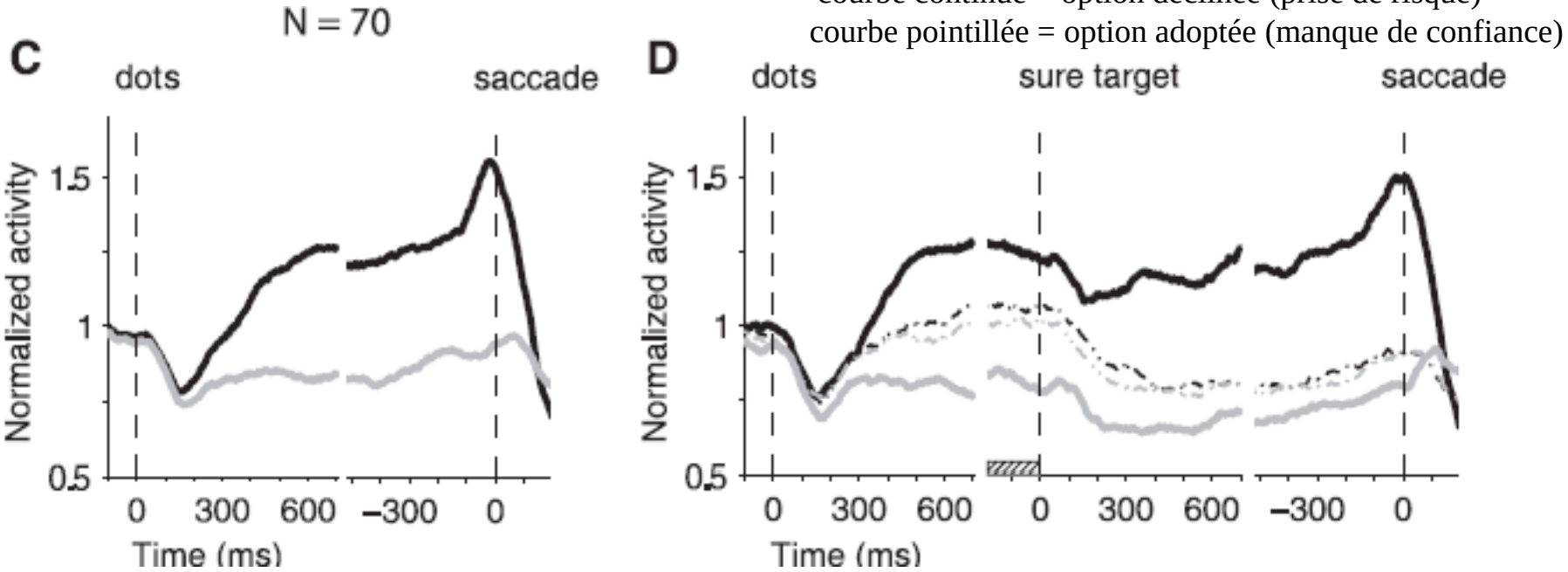
- L'estimation de l'incertitude (réponse de type II, à gauche), et la performance objective (réponse de type I, à droite) sont en image miroir.
-Surtout, la performance objective est meilleure quand l'animal dispose de l'option de refuser de répondre (courbe continue) que quand il n'en dispose pas (pointillés).
- Cela signifie que, à stimulus identique, il a sélectivement écarté les essais où il se jugeait (correctement!) incapable de répondre correctement.

Les enregistrements neuronaux sont modulés par la confiance de l'animal

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324(5928), 759-764.*

option de refuser non-disponible

option de refuser disponible:
courbe continue = option déclinée (prise de risque)
courbe pointillée = option adoptée (manque de confiance)



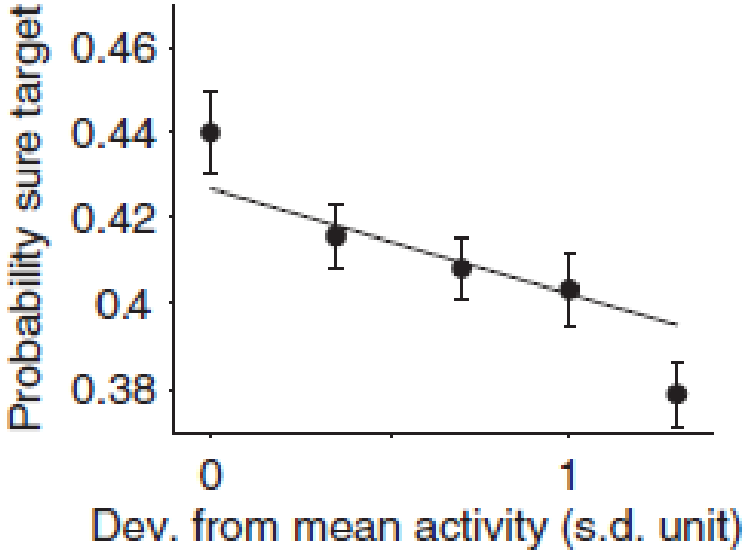
La décharge des neurones dont le champ récepteur inclut l'une des réponses primaires reflète

- la décision à venir (graphe de gauche)
- mais aussi la confiance accordée à cette réponse (graphe de droite)
- et ce, avant même que l'animal sache si l'option par défaut sera ou non proposée.

Un contrôle important: lorsque l'animal opte pour le refus de répondre, la variance des décharges neuronales est faible, ce qui rejette l'hypothèse d'une mixture d'états hauts et bas.

Les enregistrements neuronaux sont modulés par la confiance de l'animal

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324(5928), 759-764.*

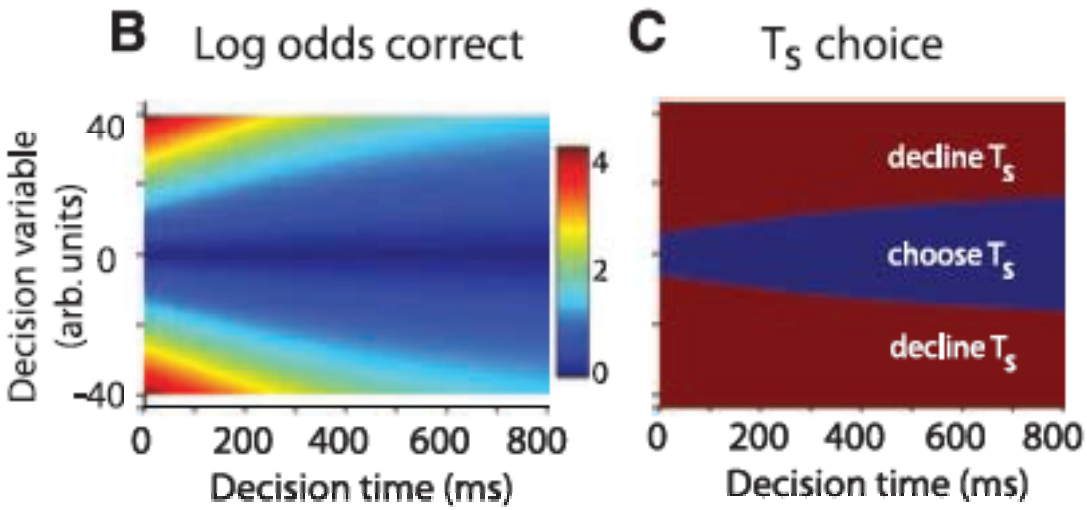


Les fluctuations d'essai en essai des décharges neuronales, mesurées juste avant la présentation de l'option de refuser de répondre, prédisent le choix de cette option.
- y compris lorsque l'on analyse des essais rigoureusement identiques sur le plan du stimulus.

La vitesse d'augmentation des décharges pendant la présentation du stimulus a également une contribution indépendante au choix de l'option de refuser de répondre.

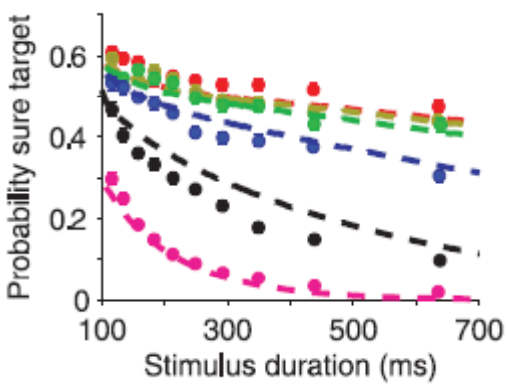
Un modèle théorique de la confiance

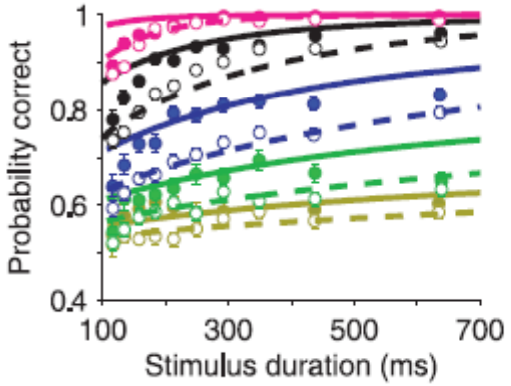
Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324(5928), 759-764.*

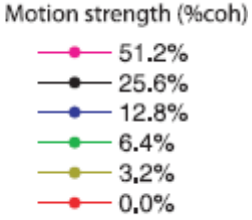


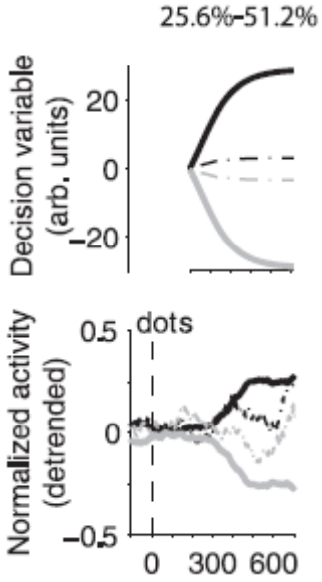
L'évolution temporelle de la variable de décision prédit assez fortement la réussite ou l'échec. Hypothèse: l'animal applique un critère de second ordre à cette variable interne, mesurée au moment de la décision.

Résultat: Cette simple hypothèse, qui généralise le modèle de la théorie de la détection du signal de second ordre, suffit à modéliser le comportement et les décharges neuronales



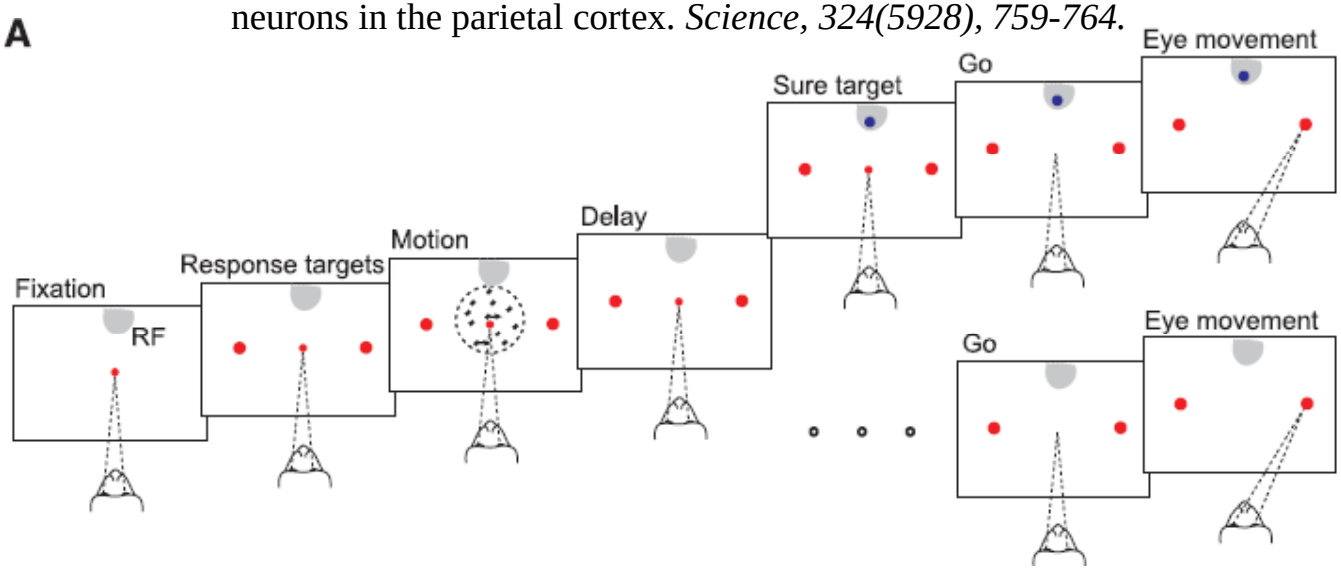


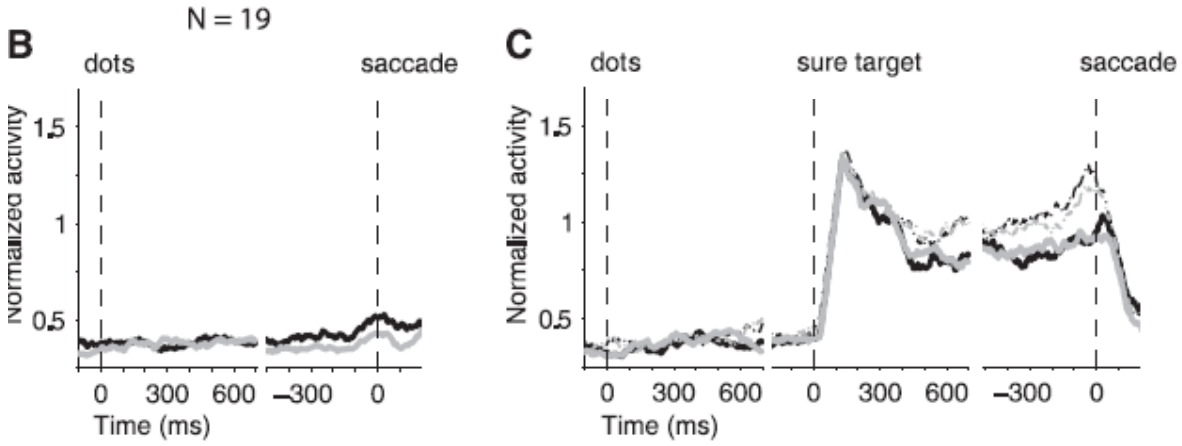


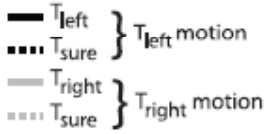


Activité neuronale liée à la décision de second ordre

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324(5928), 759-764.*





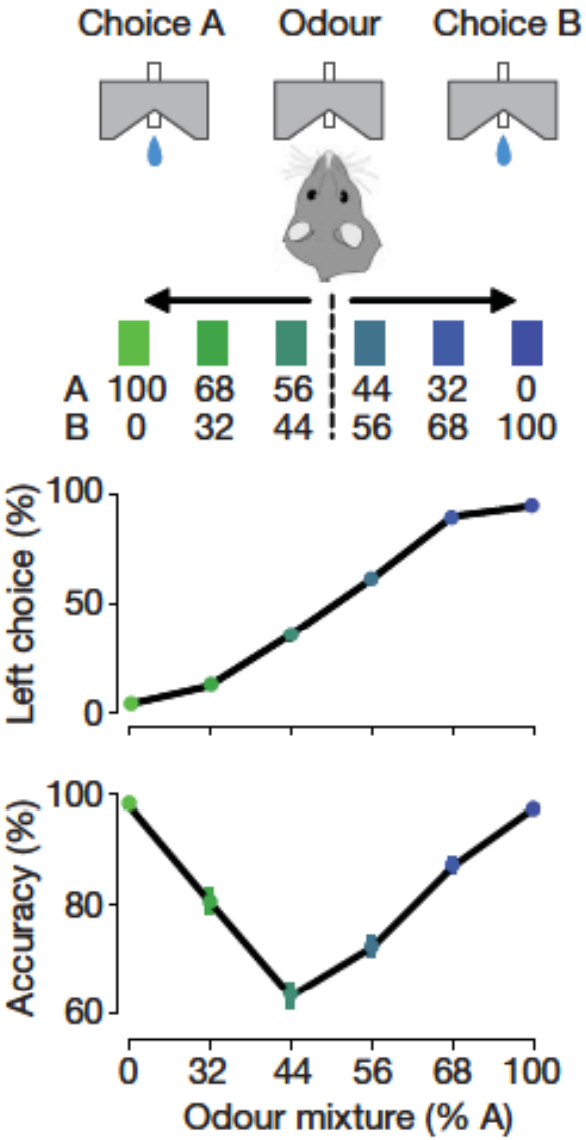


- Lorsque le champ récepteur du neurone coïncide avec la position de l'option II, on voit l'évolution du jugement de confiance.
-La décharge neuronale commence après la présentation de l'option II.

Il s'agit donc bien d'une réponse métacognitive, et non pas d'un jugement à trois alternatives.
Conclusion: le singe dispose d'un mécanisme efficace d'évaluation de ses décisions (pas nécessairement localisé dans le cortex pariétal)

Représentation neuronale de la confiance chez le rat

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455(7210), 227-231.*



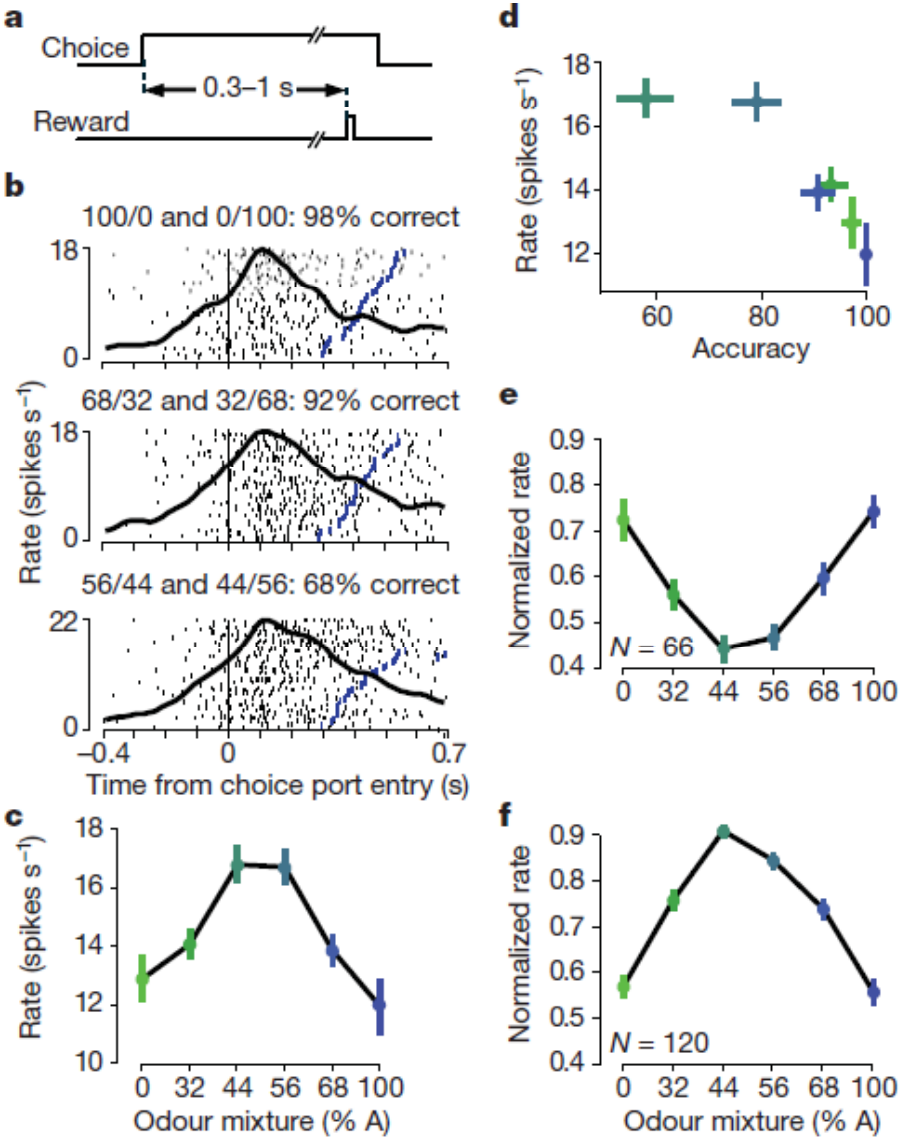
Tâche de discrimination de deux odeurs, présentées sous forme d'un mélange de proportions variables.

La récompense survient après un délai variable de 0.3 à 2 secondes.

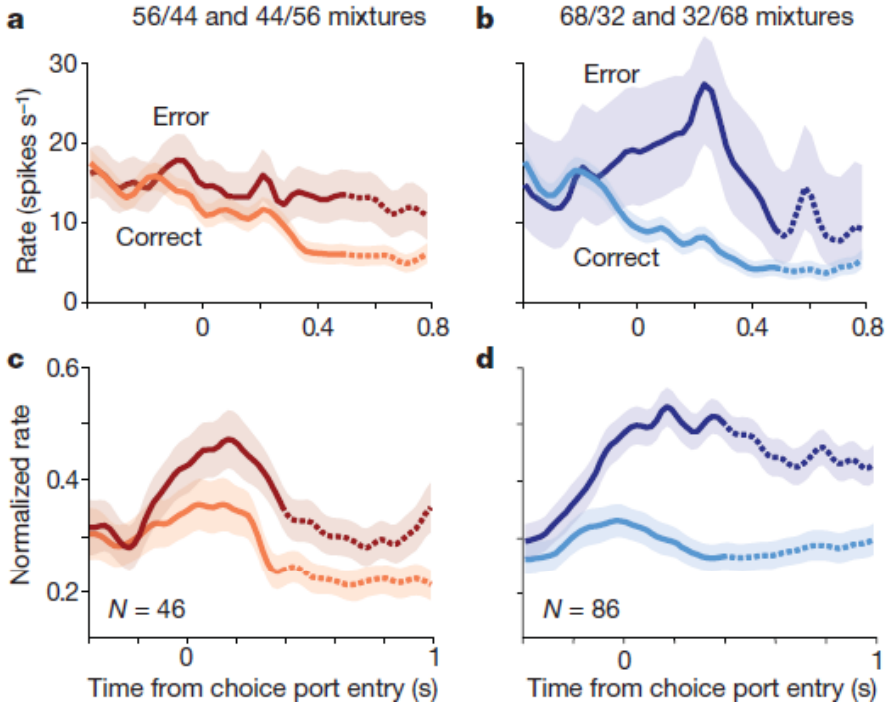
Enregistrements neuronaux dans le cortex préfrontal orbitaire, une région fréquemment associée à l'anticipation de la récompense.

Représentation neuronale de la confiance chez le rat

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455(7210), 227-231.*



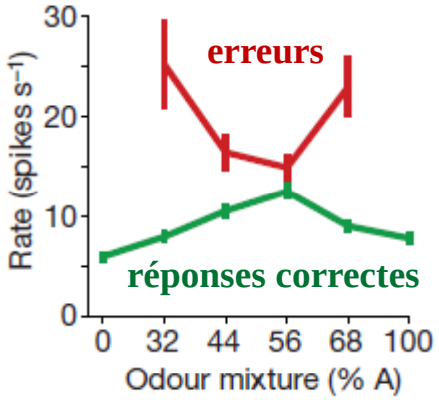
- Les réponses neuronales corrèlent, positivement ou négativement, avec la probabilité d'obtenir la récompense.
-Certaines cellules déchargent en anticipation des réponses correctes, d'autres en anticipation des réponses incorrectes.
- et ce, même à stimulus constant:

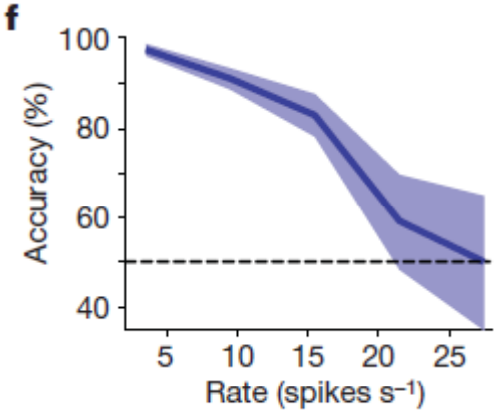


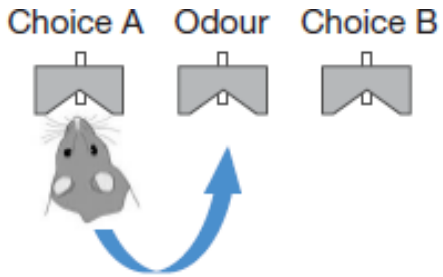
Représentation neuronale de la confiance chez le rat

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455(7210), 227-231.*

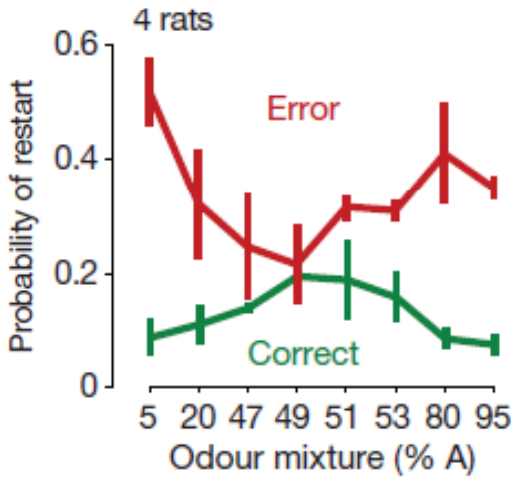
-Les réponses neuronales ne peuvent pas s'expliquer par une association entre stimuli et récompenses (même en tenant compte de l'histoire récente de ces associations).
- Par contre, elles s'expliquent bien dans la théorie de la détection du signal de type II, où le degré de confiance est déterminé par la distance entre la variable de décision et le critère de décision de type I.







-Les rats sont capables d'utiliser cette information:
Lorsqu'on leur donne la possibilité de redémarrer un essai sans attendre la récompense, en revenant au centre, ils le font d'autant plus souvent qu'ils se sont trompés, avec une courbe très similaire aux réponses neuronales.



Il pourrait donc s'agir d'une anticipation de la récompense, mais fondée sur l'analyse de la qualité de la décision et pas seulement sur le stimulus présenté → une authentique métacognition chez le rat

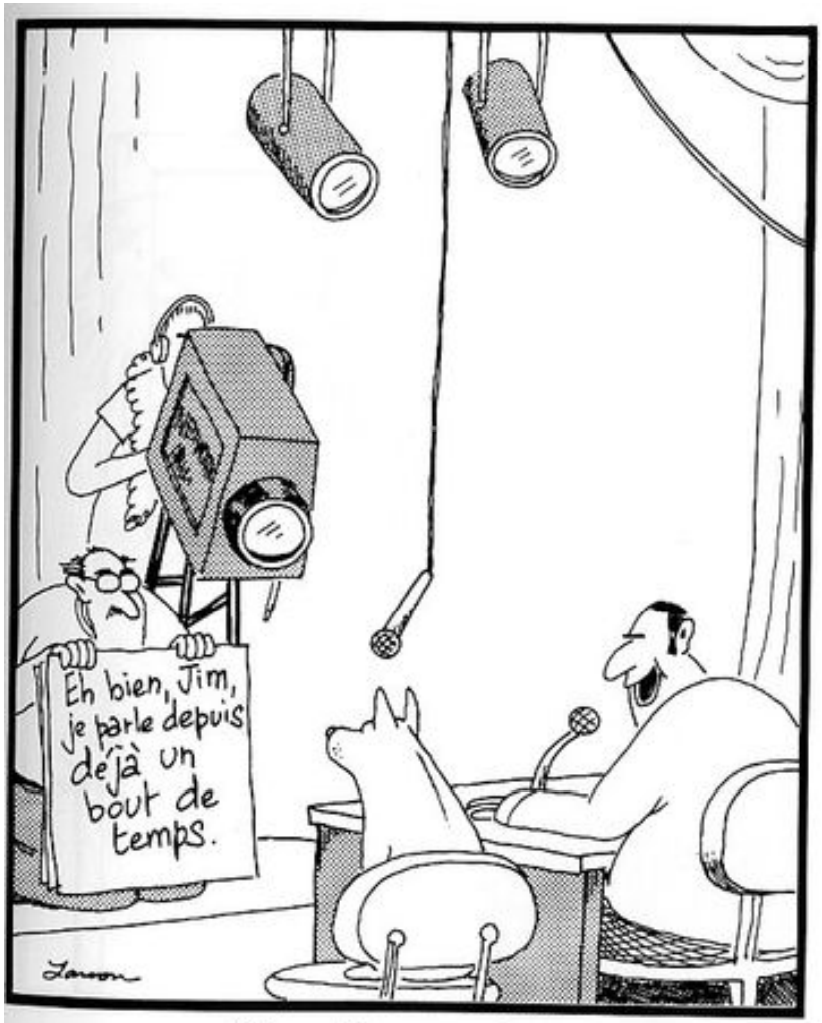
Conclusions

L’estimation de l’incertitude semble faire partie intégrante de la décision.
A ce titre, elle est présente chez de nombreuses espèces animales.

Plus impressionnant est le fait que ces animaux parviennent à utiliser leur estimation de l’incertitude pour modifier leur comportement.
Il s’agit véritablement d’un jugement de second ordre ou métacognitif (mais pas nécessairement conscient).

Cependant, les expériences de laboratoires posent toujours la question de l’entraînement intensif de l’animal…

Une approche éthologique reste à mener afin de vérifier si de tels jugements sont utilisés en milieu naturel.



"Alors, dites-nous, Rex...
Depuis combien de temps êtes-vous un chien parlant ?"