Tronc commun « Sciences Cognitives » de l'Ecole Polytechnique

Stanislas Dehaene

Professeur au Collège de France
Chaire de Psychologie Cognitive Expérimentale

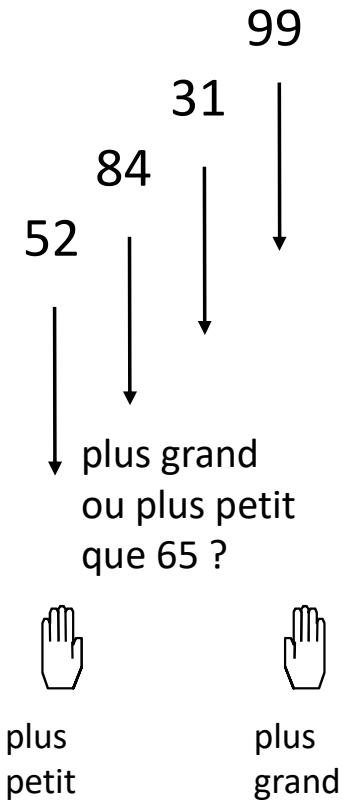
Cours n°2

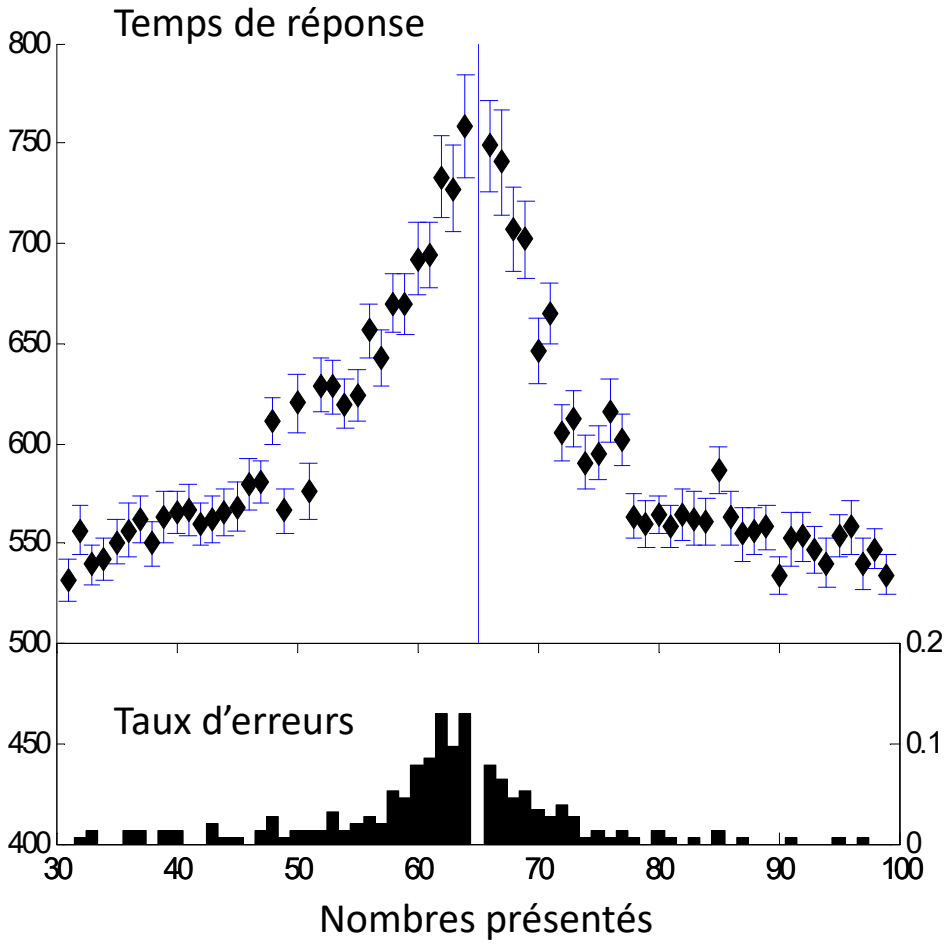
Le cerveau bayésien :
Le rôle des statistiques dans la perception, la prise de décision et l'apprentissage

Un exemple de prise de décision

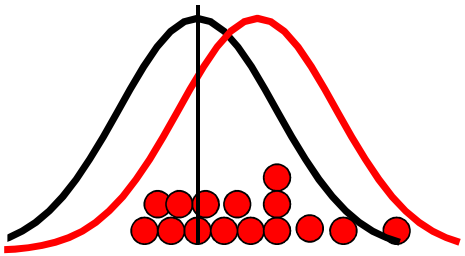
L'effet de distance en comparaison de nombres

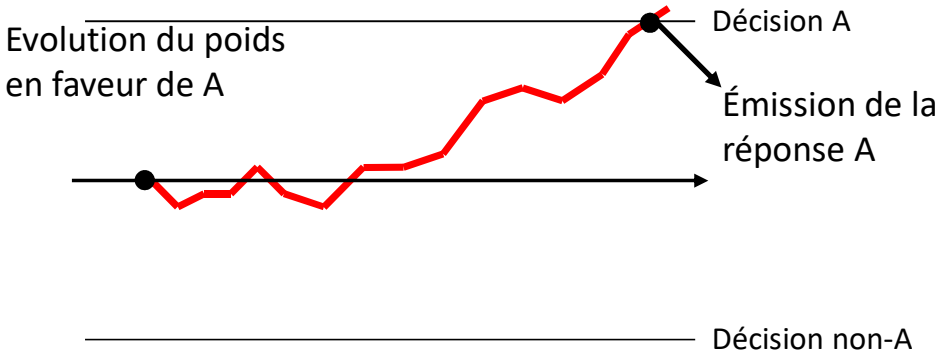
découvert par Moyer et Landauer en 1967





Analyse mathématique:
La prise de décision comme une marche aléatoire





Justification par l'inférence Bayésienne:
L'algorithme de Wald, découvert indépendamment par Alan Turing à Bletchley Park

Decision rule (BF) — Sequential sampling rule — Bayes' rule — Independence property

$$\frac{p(\mathrm{L}|s)}{p(\mathrm{R}|s)} = \frac{p(\mathrm{L}|s_1, s_2, \dots, s_n)}{p(\mathrm{R}|s_1, s_2, \dots, s_n)} = \frac{p(s_1, s_2, \dots, s_n|\mathrm{L}) \cdot p(\mathrm{L})}{p(s_1, s_2, \dots, s_n|\mathrm{R}) \cdot p(\mathrm{R})} = \frac{p(\mathrm{L})}{p(\mathrm{R})} \cdot \frac{\prod_{t=1}^{n} p(s_t|\mathrm{L})}{\prod_{t=1}^{n} p(s_t|\mathrm{R})}$$

$$\frac{p(\mathrm{L})}{p(\mathrm{R})} \cdot \prod_{t=1}^{n} \frac{p(s_t|\mathrm{L})}{p(s_t|\mathrm{R})}$$

Log-scale

$$v(t) = \log \frac{p(\mathrm{L})}{p(\mathrm{R})} + \sum_{t=1}^{n} \log \frac{p(s_t|\mathrm{L})}{p(s_t|\mathrm{R})}$$

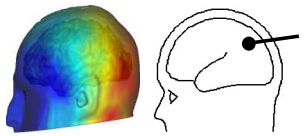
Décomposition d'une tâche cognitive

Reconnaissance des symboles	***Conversion en quantité et décision***	***Programmation de la réponse***
Cinquante-huit	58	Main droite

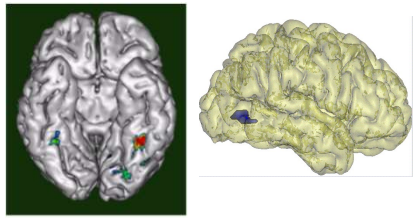
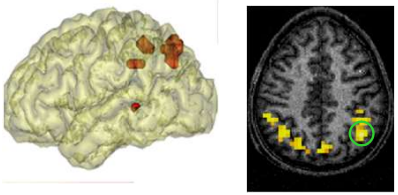
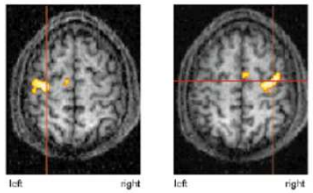
Signature comportementale

Effet de notation	Effet de distance	Effets de latéralisation et de complexité motrice

Déroulement temporel en potentiels évoqués

110-170 ms	190-300 ms	250-450 ms
		

Localisation des circuits par IRM fonctionnelle

Région occipito-temporale ventrale	Région intrapariétale	Régions prémotrices et motrices
		

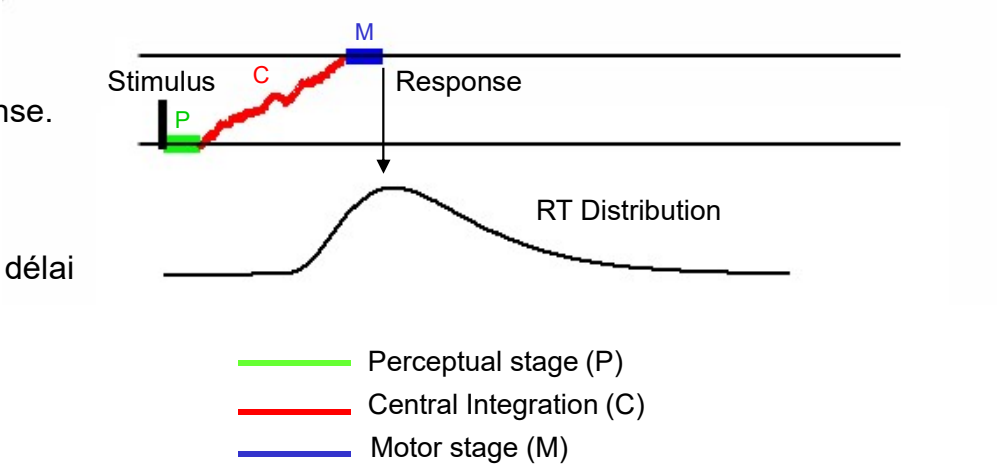
L. Cohen, L. Naccache, P. Pinel

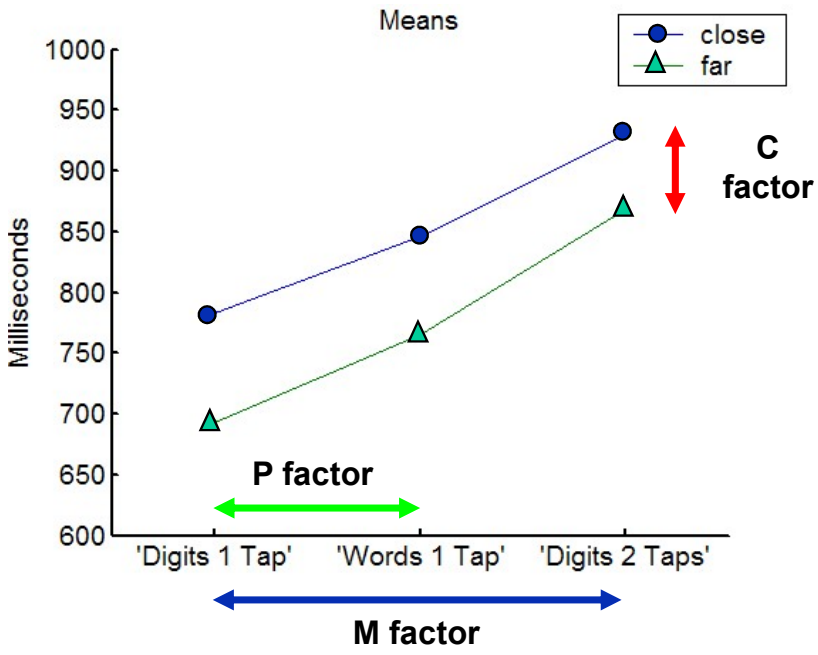
Ce modèle fait des prédictions spécifiques :

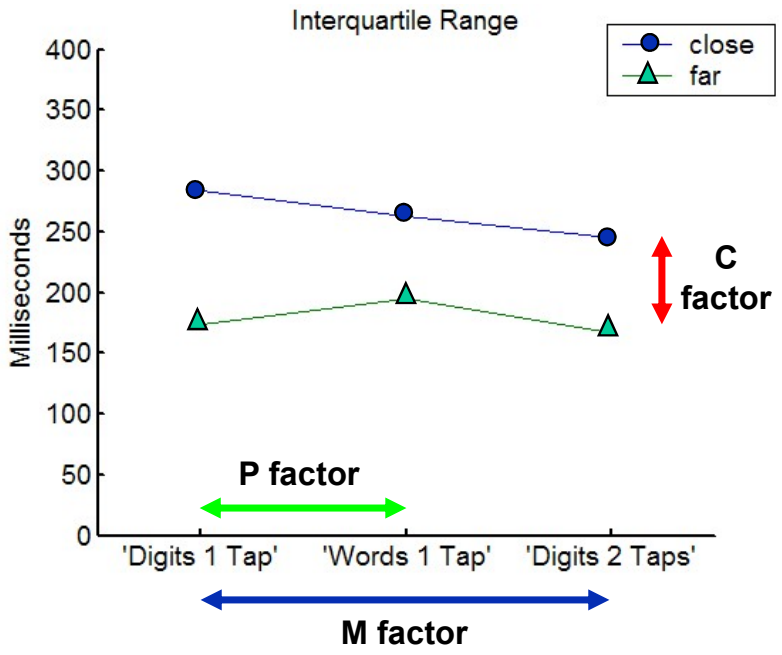
1. Additivité des trois effets sur les moyennes des temps de réponse.

2. Effets différents sur la **variabilité** des temps de réponse

- Les facteurs qui affectent les étapes P ou M doivent ajouter un délai fixe (pas d'effet sur la variabilité)**:**
  - notation (chiffres ou mots)
  - Complexité de la réponse (un ou deux clics)
- Les facteurs qui affectent C doivent **augmenter la variabilité**
  - Distance numérique



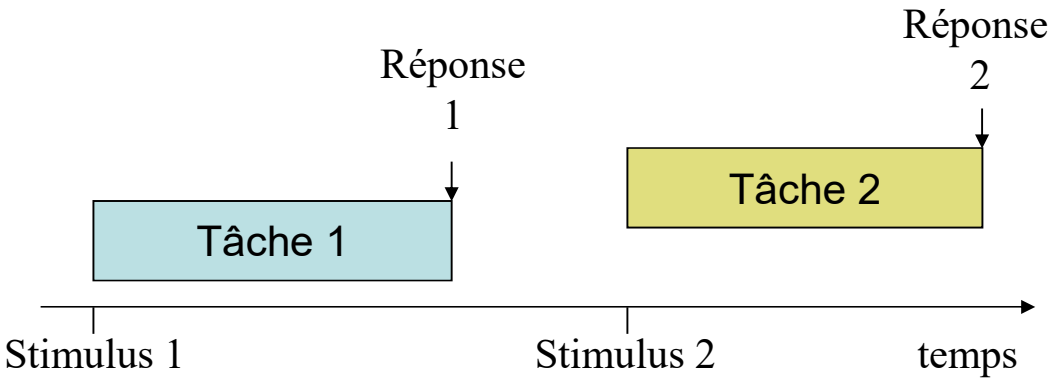




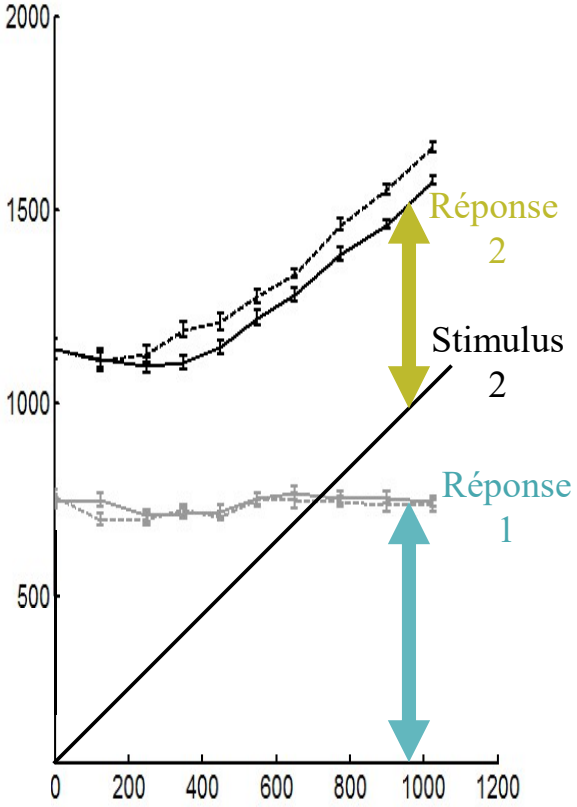
Sigman & Dehaene, PLOS: Biology, 2004

Collisions mentales dans l'exécution simultanée de deux tâches

La « période psychologique réfractaire »



Moment où survient la réponse

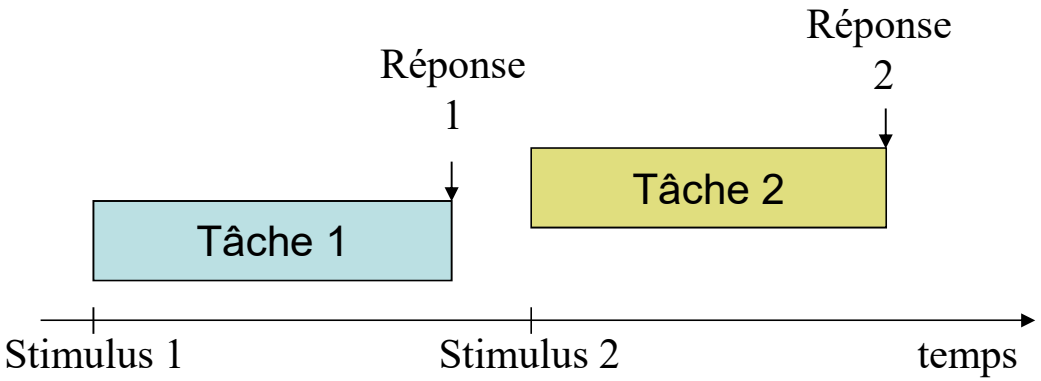


Intervalle entre les stimuli

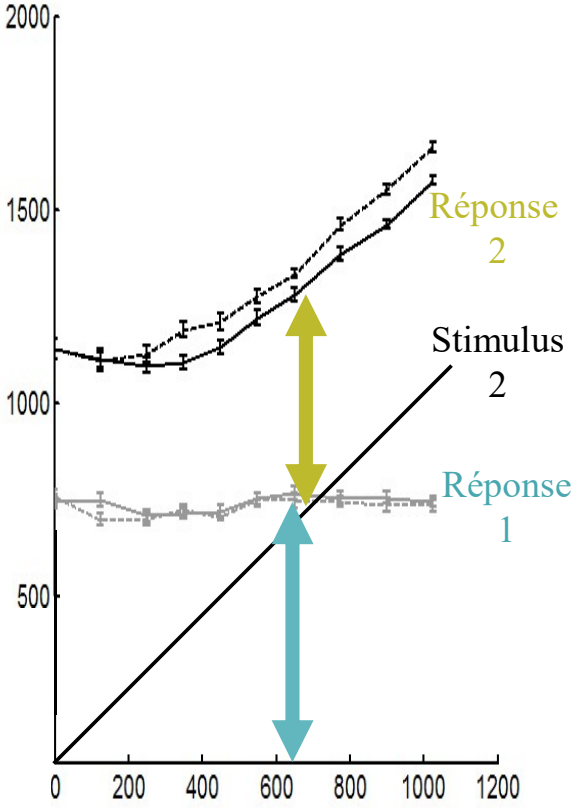
M. Sigman

Collisions mentales dans l'exécution simultanée de deux tâches

La « période psychologique réfractaire »



Moment où survient la réponse

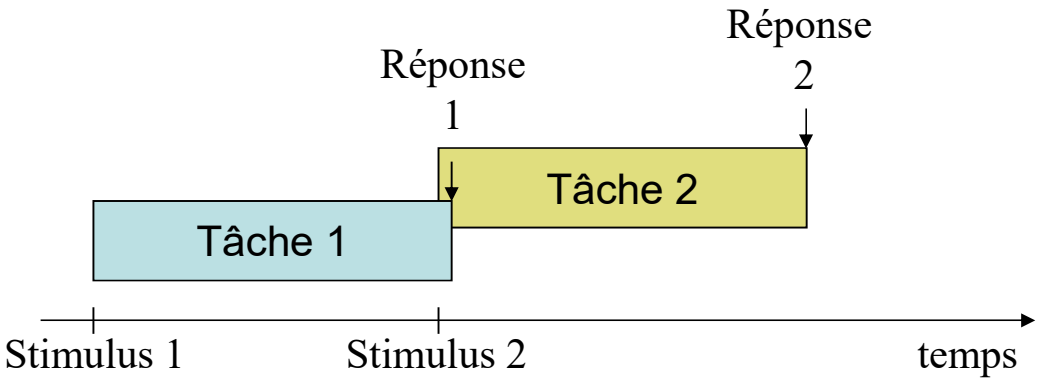


Intervalle entre les stimuli

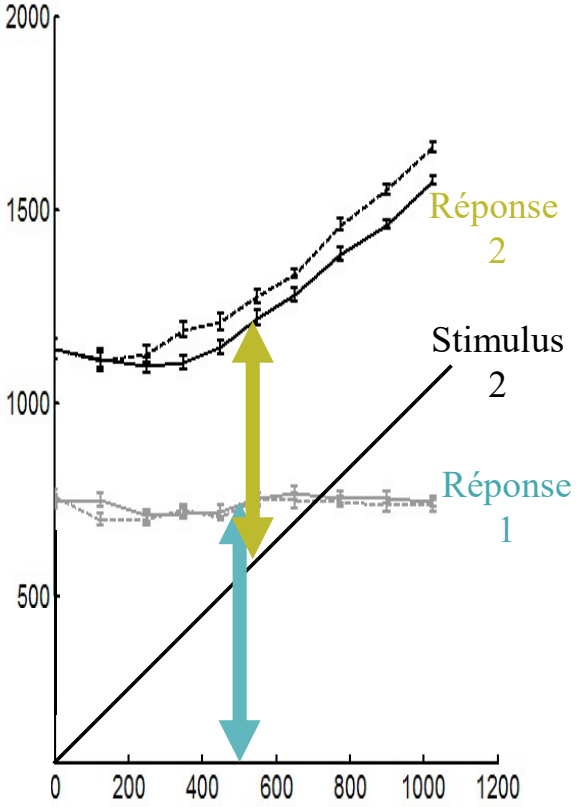
M. Sigman

Collisions mentales dans l'exécution simultanée de deux tâches

La « période psychologique réfractaire »



Moment où survient la réponse

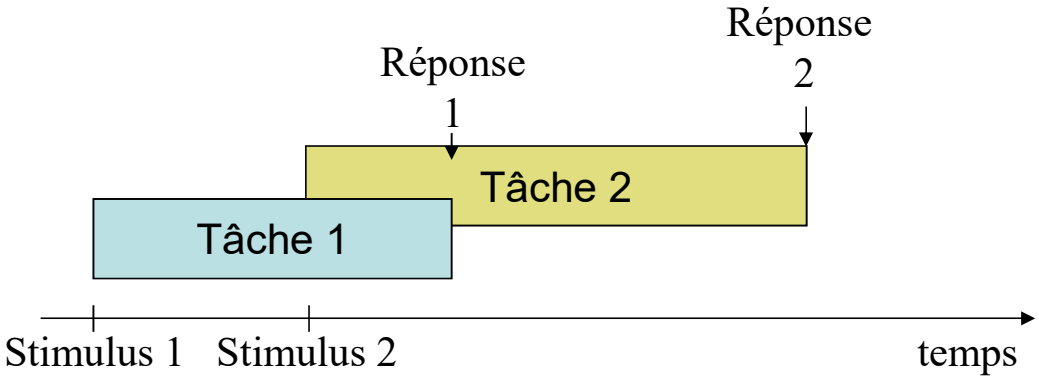


Intervalle entre les stimuli

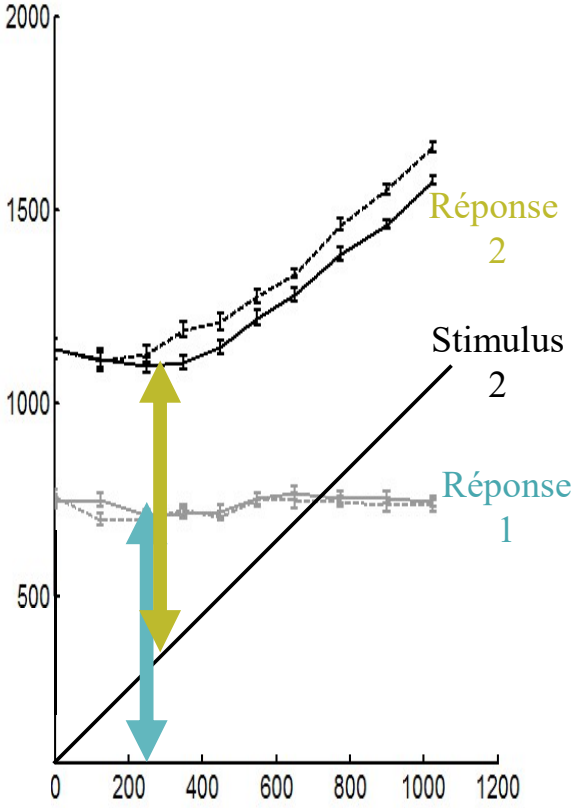
M. Sigman

Collisions mentales dans l'exécution simultanée de deux tâches

La « période psychologique réfractaire »



Moment où survient la réponse

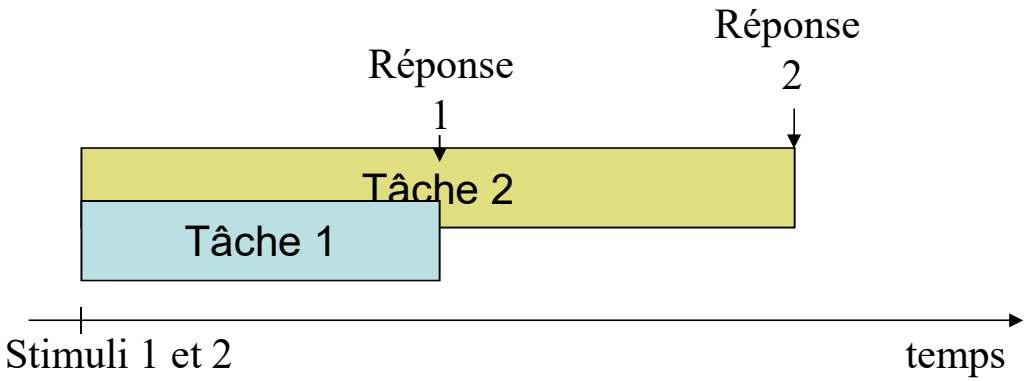


Intervalle entre les stimuli

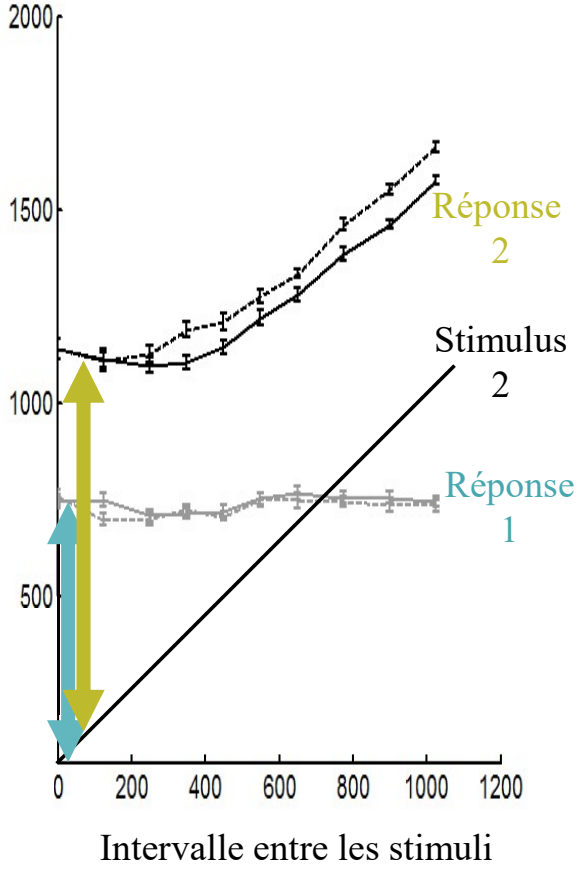
M. Sigman

Collisions mentales dans l'exécution simultanée de deux tâches

La « période psychologique réfractaire »



Moment où survient la réponse

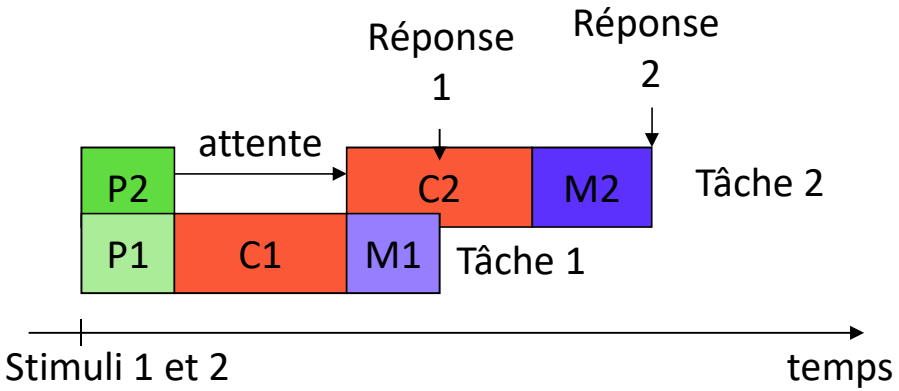


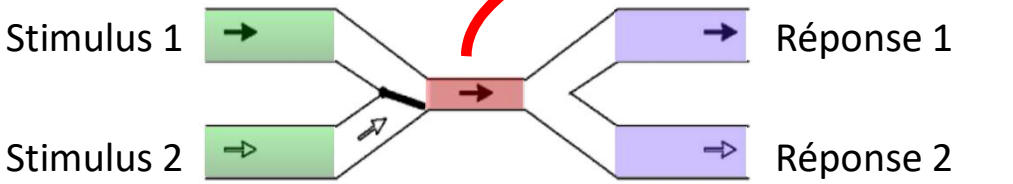
Intervalle entre les stimuli

M. Sigman

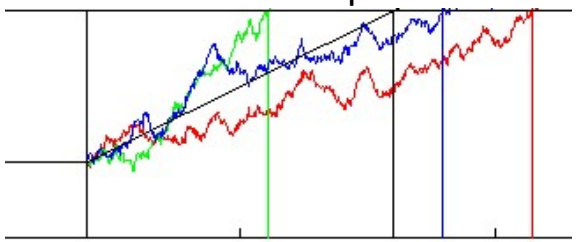
Le goulot d'étranglement central de Pashler

- Les deux tâches ne sont pas exécutées en parallèle, mais partiellement en série.
- Seule une étape « centrale » est ralentie durant la collision
- La durée de l'étape P2 peut être « absorbée » durant le temps d'attente.



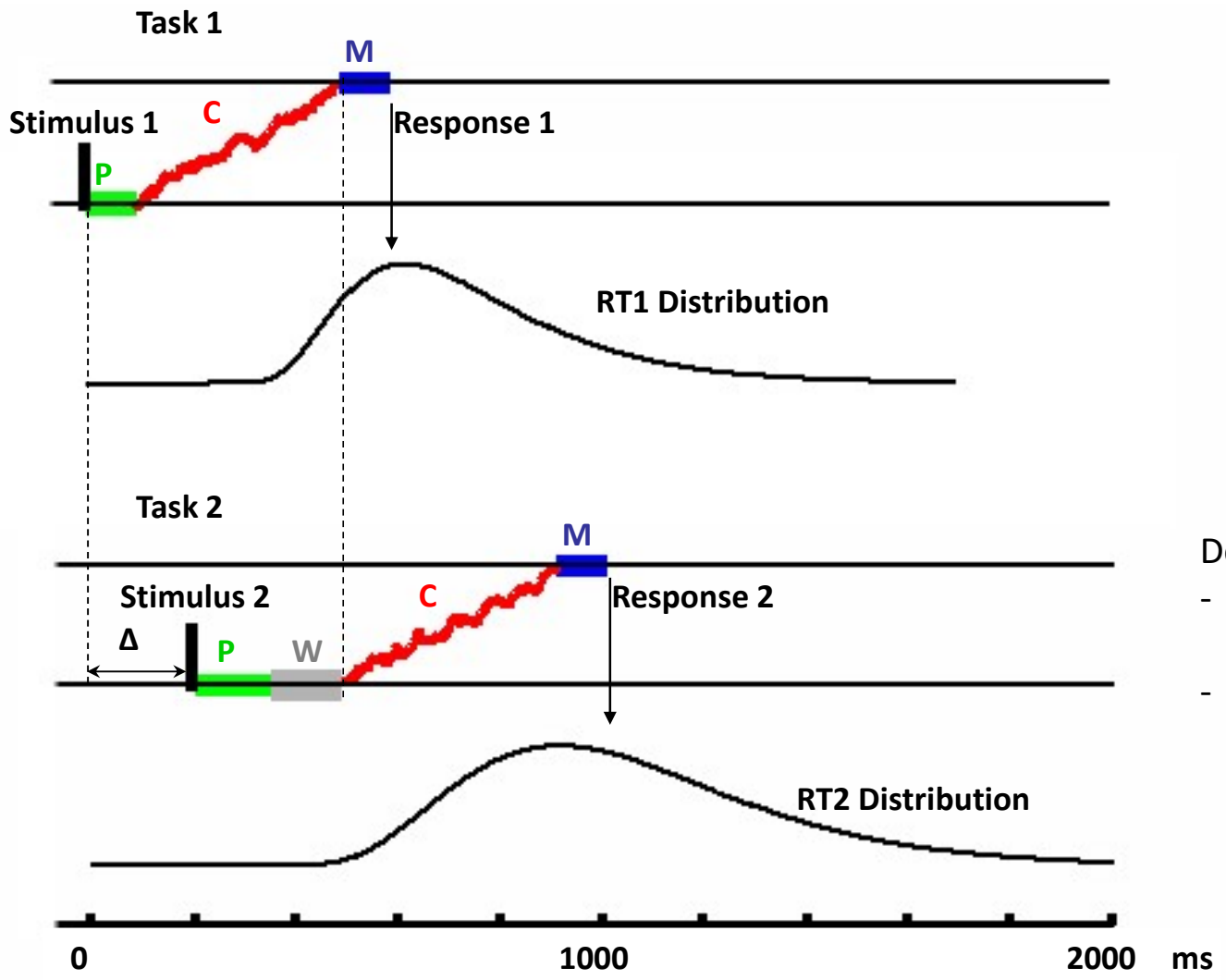


Cette étape centrale correspond à la prise de décision stochastique



Comment se comporte la variabilité en cas de double tâche?

Sigman & Dehaene, PLOS Biology 2005

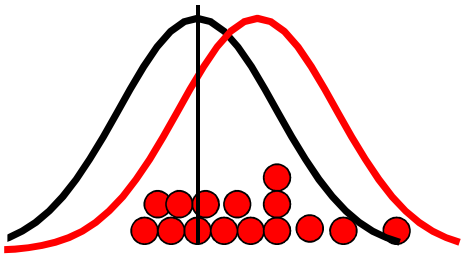


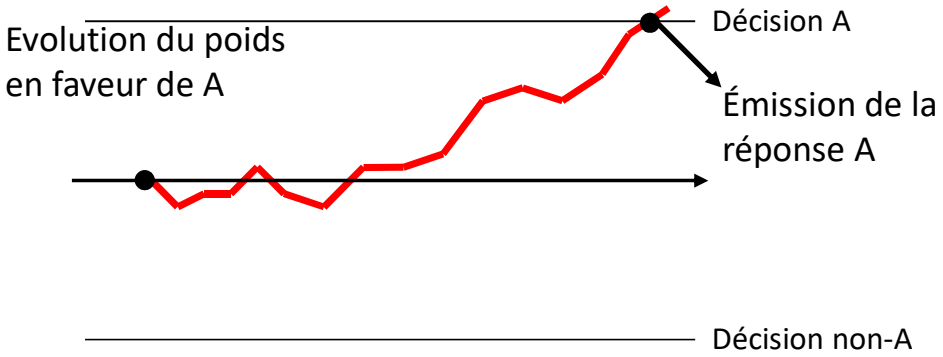
Perceptual stage (P)
Wait period (W)
Central Integration (C)
Motor stage (M)

Deux prédictions:

- Le temps d'attente dépend de la durée du temps de réponse 1
- La variabilité du temps de réponse 2 intègre, en partie, la variance du temps de réponse 1.

Chaque décision est une inférence





Le cerveau infère quelle est la réponse la plus probable, étant données les entrées sensorielles, en utilisant la règle de Bayes.

Decision rule (BF) — Sequential sampling rule — Bayes' rule — Independence property

$$\frac{p(\mathrm{L}|s)}{p(\mathrm{R}|s)} = \frac{p(\mathrm{L}|s_1, s_2, \ldots, s_n)}{p(\mathrm{R}|s_1, s_2, \ldots, s_n)} = \frac{p(s_1, s_2, \ldots, s_n|\mathrm{L}) \cdot p(\mathrm{L})}{p(s_1, s_2, \ldots, s_n|\mathrm{R}) \cdot p(\mathrm{R})} = \frac{p(\mathrm{L})}{p(\mathrm{R})} \cdot \frac{\prod_{t=1}^{n} p(s_t|\mathrm{L})}{\prod_{t=1}^{n} p(s_t|\mathrm{R})}$$

$$v(t) = \log\frac{p(\mathrm{L})}{p(\mathrm{R})} + \sum_{t=1}^{n} \log\frac{p(s_t|\mathrm{L})}{p(s_t|\mathrm{R})}$$

Log-scale ←

$$\frac{p(\mathrm{L})}{p(\mathrm{R})} \cdot \prod_{t=1}^{n} \frac{p(s_t|\mathrm{L})}{p(s_t|\mathrm{R})}$$

La perception est également une inférence

La perception est également une inférence

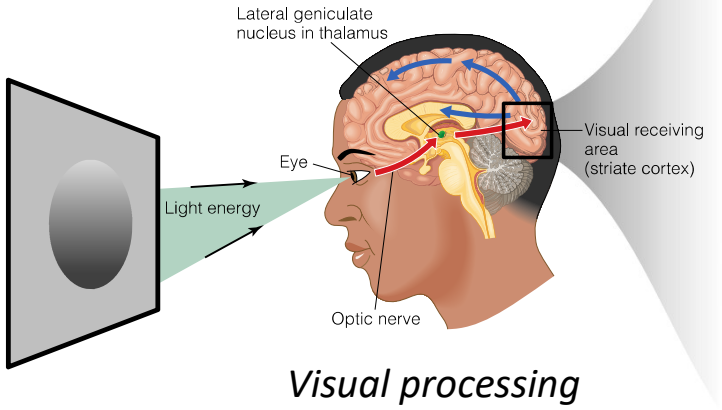
Le cerveau reçoit des informations sensorielles, et en **infère** l'état le plus probable du monde.

Selon Helmholtz (1851), les entrées sensorielles (*Perception*) se combinent avec les attentes (*Vorstellung*) pour engendrer l'experience consciente (*Anschauung*).

L'analyse Bayésienne constitue un modèle normatif de cette idée:

Hermann Von Helmhotz
(1821 – 1894)

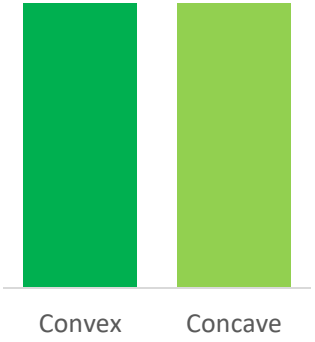
Visual stimulation **Information processing**



Visual processing

Likelihood $p(d|h)$

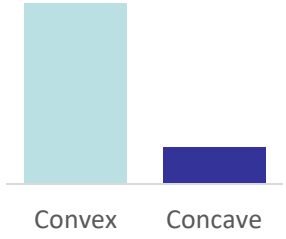
How probable is the retinal image if the hypothesis were true?



×

Prior $p(h)$

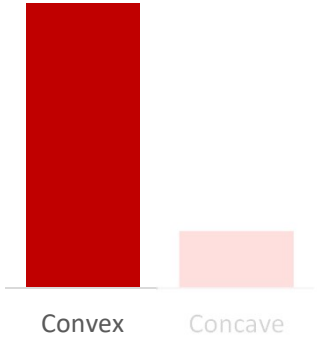
How much do you except the hypothesis based on your past experiences?



∝

Posterior $p(h|d)$

The most probable hypothesis becomes your percept!



Raisonnement probabiliste « en avant » (classique) ou « en arrière » (bayésien)

La théorie des probabilités, telle qu'enseignée à l'école, est surtout utilisée pour calculer la probabilité d'une observation, étant donné certaines hypothèses sur l'état du monde.

Par exemple, soit une urne contenant 3 boules noires et 7 boules blanches. Quelle est la probabilité, lors de deux tirages sans remplacement, de tirer deux boules noires?

$p(H \,\&\, N_1 \,\&\, N_2) = p(H)\, p(N_1|H)\, p(N_2|N_1 \,\&\, H) = (3/10) \times (2/9) = 1/15$

Mais, les données d'observation D et les hypothèses H jouent des rôles strictement symétriques. Rien n'empêche d'utiliser les équations pour inverser le procédé:

Etant donnée l'observation D, quelle est la probabilité de l'hypothèse H?

Application de la règle fondamentale: $p(H \,\&\, D) = p(D|H)\, p(H) = p(H|D)\, p(D)$

D'où $p(H|D) = p(D|H)\, p(H) / p(D)$ ou $p(H|D) \;\alpha\; p(D|H)\, p(H)$

p(H) = *probabilité « a priori » de H* (*prior* en anglais) (mais pas dans le sens Kantien d'indépendant de l'expérience; elle peut résulter d'expériences antérieures)

p(H|D) = *probabilité « a posteriori » de H* (pas nécessairement au sens temporel, mais au sens de la déduction logique, après avoir observé D)

p(D|H), considéré comme une fonction de H, est la *vraisemblance de H*

Un exemple qualitatif de raisonnement Bayésien

Mon fils Olivier tousse. Trois hypothèses:

h1=il a la grippe. h2=il a un cancer du poumon. h3=il a une gastro-entérite.

Règle fondamentale: $p(H|D) \ \alpha \ p(D|H)\ p(H)$

Pour h1 (grippe): tant la probabilité a priori que la vraisemblance sont élevées

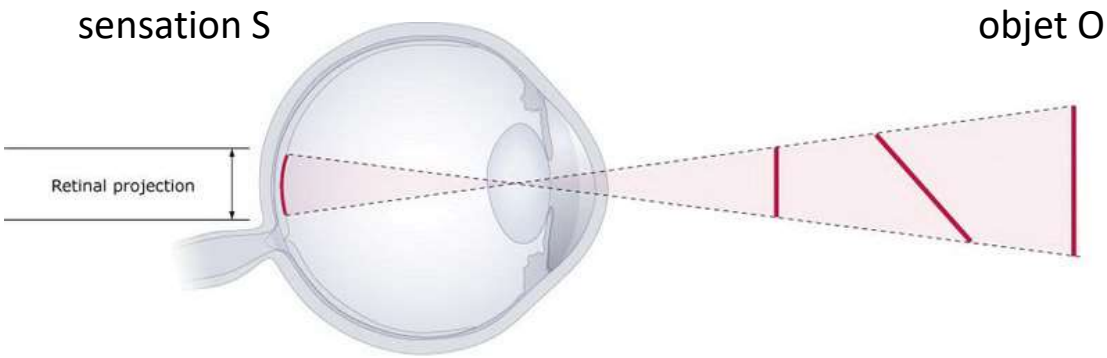
Pour h2 (cancer du poumon): la vraisemblance est élevée mais la probabilité a priori est faible

Pour h3 (gastro-entérite): la probabilité a priori est élevée, mais la vraisemblance est faible.

Ainsi h1 est l'hypothèse la plus probable a posteriori .

Nous utilisons un critère « MAP » (*maximum a posteriori* hypothesis) et non un critère de maximum de vraisemblance (ML ou *maximum likelihood*) car la probabilité *a priori* ne peut pas être négligée.

Pourquoi le cadre Bayésien s'applique-t-il bien à la perception?



Les entrées sensorielles sont *toujours* ambigües.

Par exemple, un cercle se projette comme une ellipse sur la rétine.

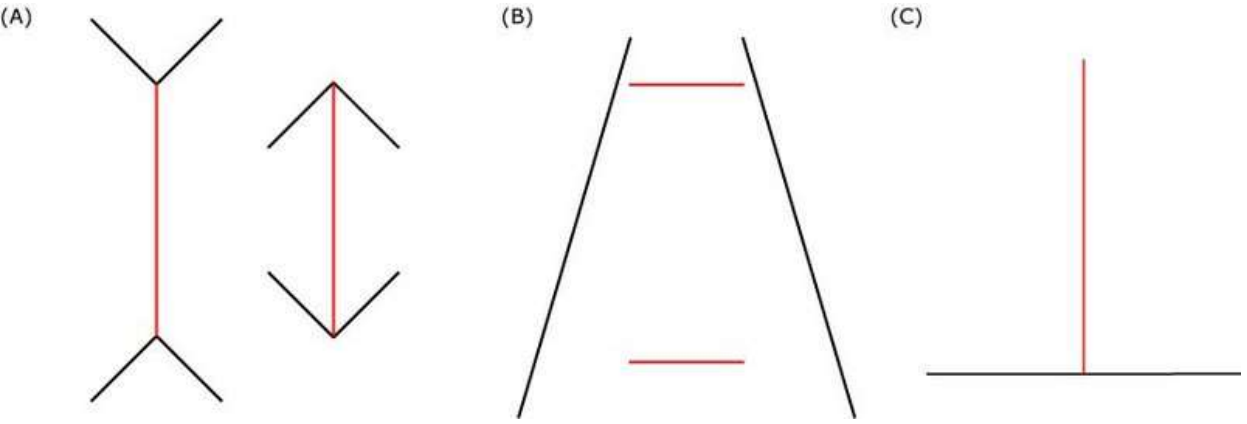
Notre système perceptif doit donc *sélectionner*, parmi une infinité de solutions possibles, celle qui est la plus *plausible*.

La théorie Bayésienne explique ce processus de choix sur la base de:
- la connaissance *a priori* des probabilités des objets dans le monde extérieur p(O)
(qui peut résulter d'une accumulation de connaissances au cours de l'apprentissage)
- la connaissance de la fonction de vraisemblance p(S|O)
(qui peut résulter d'un *modèle interne* du comportement des objets)
- l'application de la règle de Bayes: $p(O|S) \alpha p(S|O)\, p(O)$

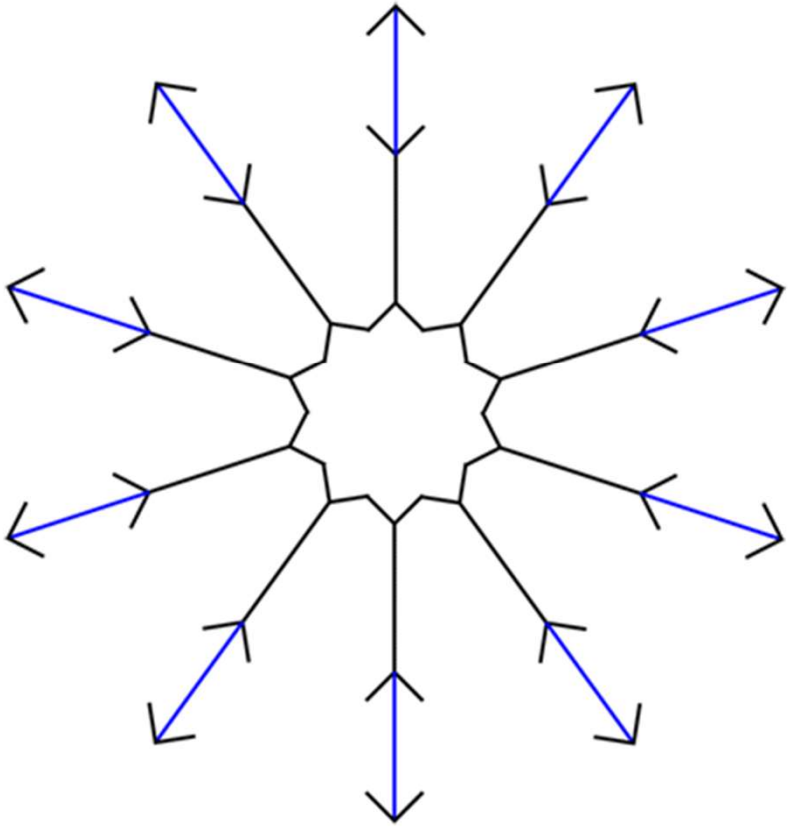
L’effet massif de la probabilité a priori:
Les illusions visuelles reflètent souvent un retour à l’interprétation la plus probable

Exemple: la perception de la taille

La taille perçue dépend de la projection rétinienne, mais aussi du contexte.



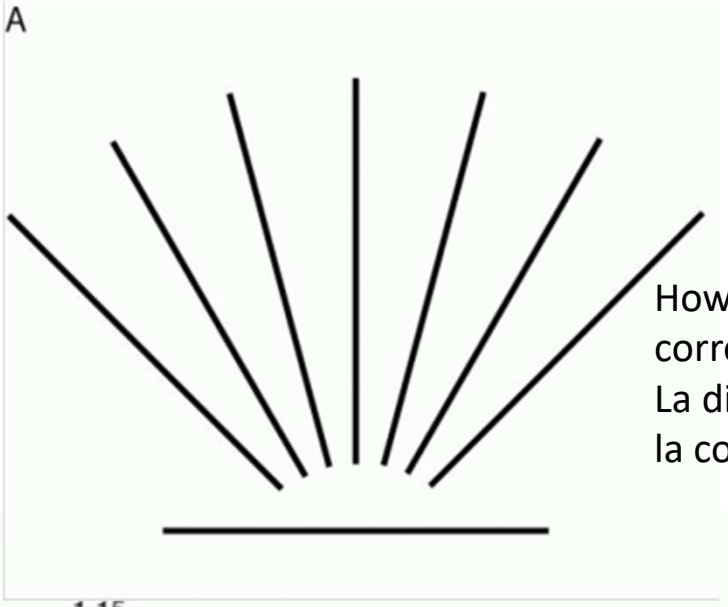
Sarcone's Dynamic Müller-Lyer Illusion

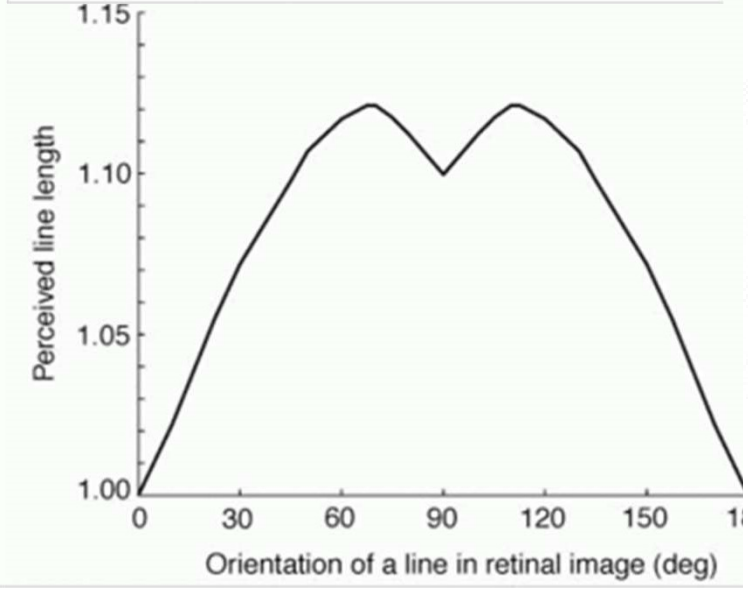


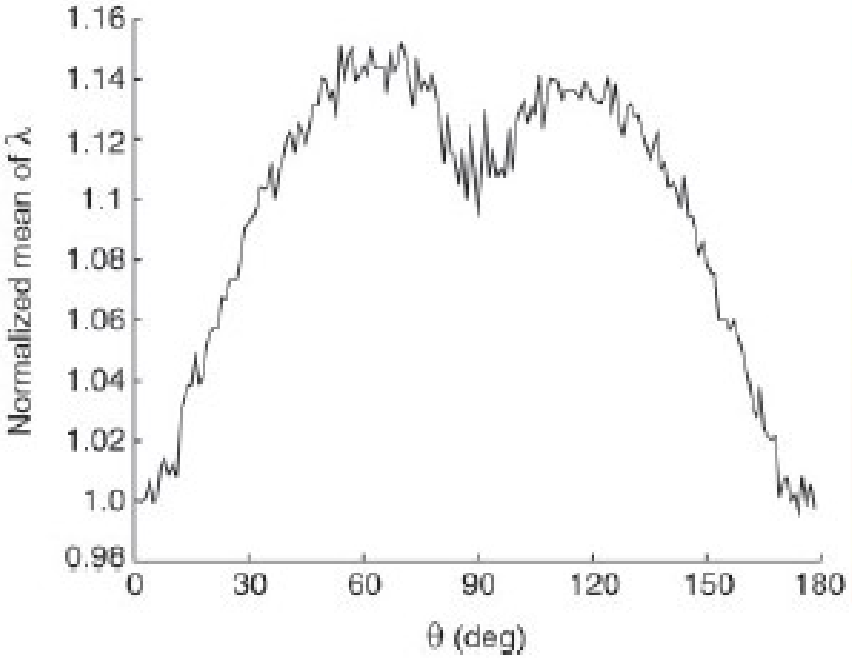
Exemple: la perception de la taille

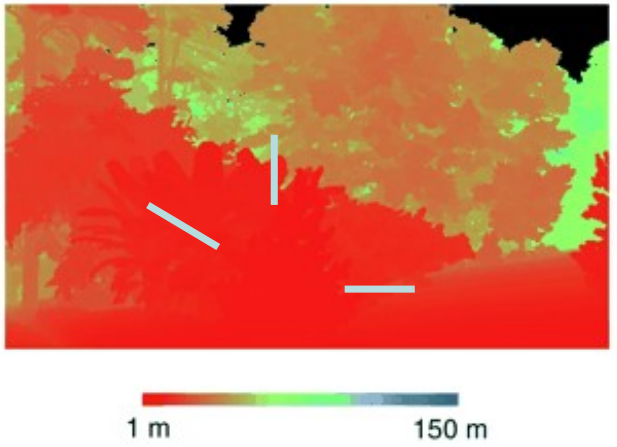
Howe, C. Q., & Purves, D. (2002). Range image statistics can explain the anomalous perception of length. *Proc Natl Acad Sci U S A, 99(20), 13184-13188.*

Howe et Purves ont mesuré, par télémétrie laser, les distances dans le monde réel correspondant à une distance fixe sur la rétine.
La distance mesurée pour un angle donné reproduit la courbe de l'illusion perceptive:

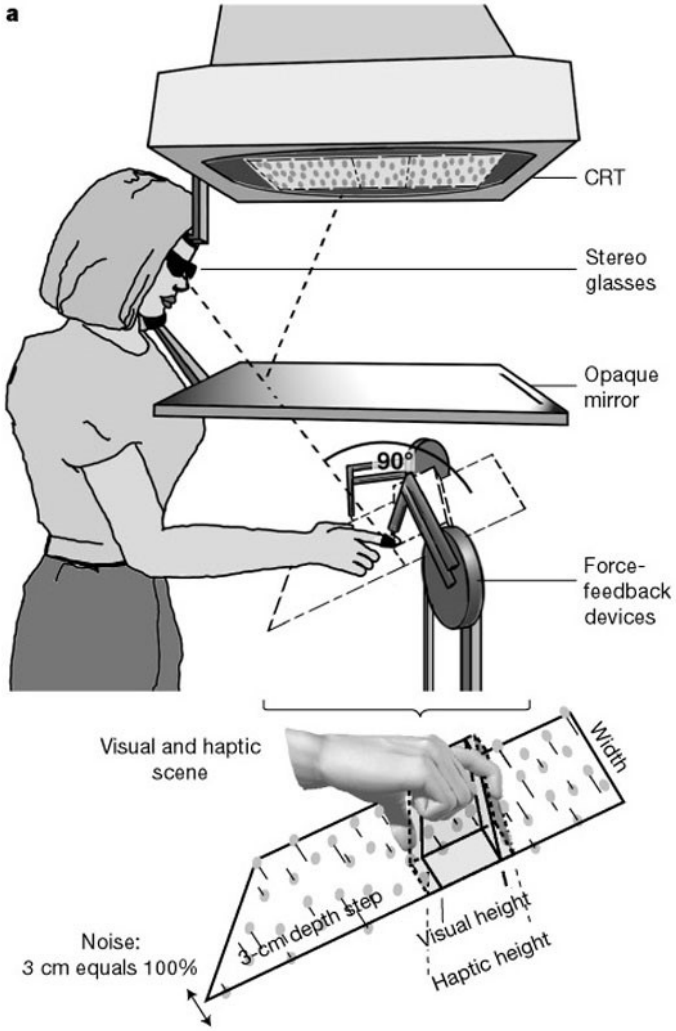








L’intégration Bayésienne de plusieurs indices sensoriels

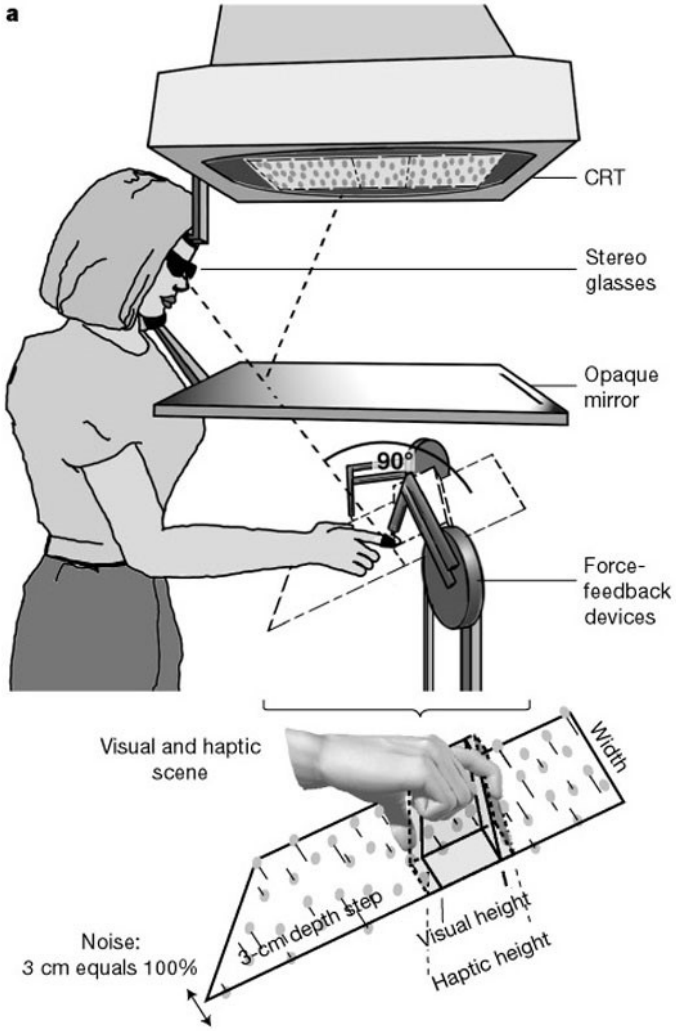


Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415(6870), 429-433.*

Imaginons que nous recevions simultanément des indices visuels et tactiles sur la taille d’un objet.

Comment combiner ces deux informations?

L'intégration Bayésienne de plusieurs indices sensoriels



Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415(6870), 429-433.*

Si ces indices sont conditionnellement indépendants, la théorie Bayésienne nous dit que la densité de probabilité de leur combinaison est le produit des deux densités.

$$P(w|t,v) = \frac{P(t,v|w)P(w)}{P(t,v)}$$
$$= \frac{P(t|w)P(v|w)P(w)}{P(t,v)}$$
$$\propto P(t|w)P(v|w)P(w)$$

L'intégration Bayésienne de plusieurs indices sensoriels

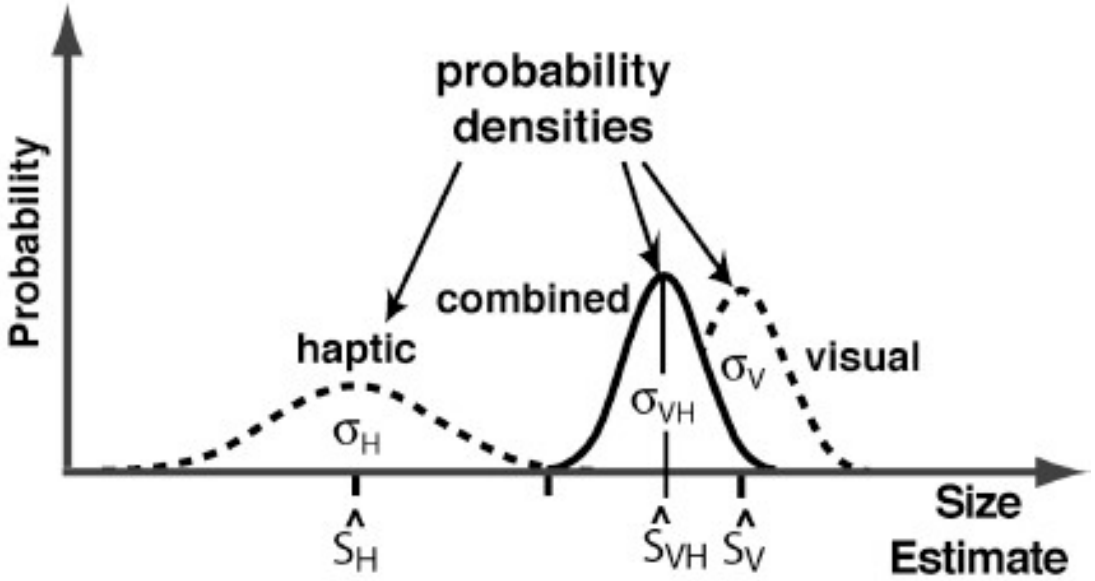
Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415(6870), 429-433.*

Le produit de deux Gaussiennes est une nouvelle Gaussienne
Le principe du maximum de vraisemblance prédit que
-la perception est une moyenne pondérée des valeurs suggérées par chaque indice
-les pondérations sont fonction de la fiabilité (*reliability*) des indices (l'inverse de la variance)
-la fiabilité totale est la somme des fiabilités (l'information de Fischer est additive pour des signaux indépendants)

$$\hat{S} = \sum_i w_i \hat{S}_i$$

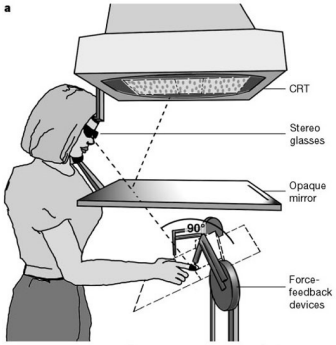
$$w_j = \frac{r_j}{\sum_i r_i}$$

$$r = \sum_i r_i$$

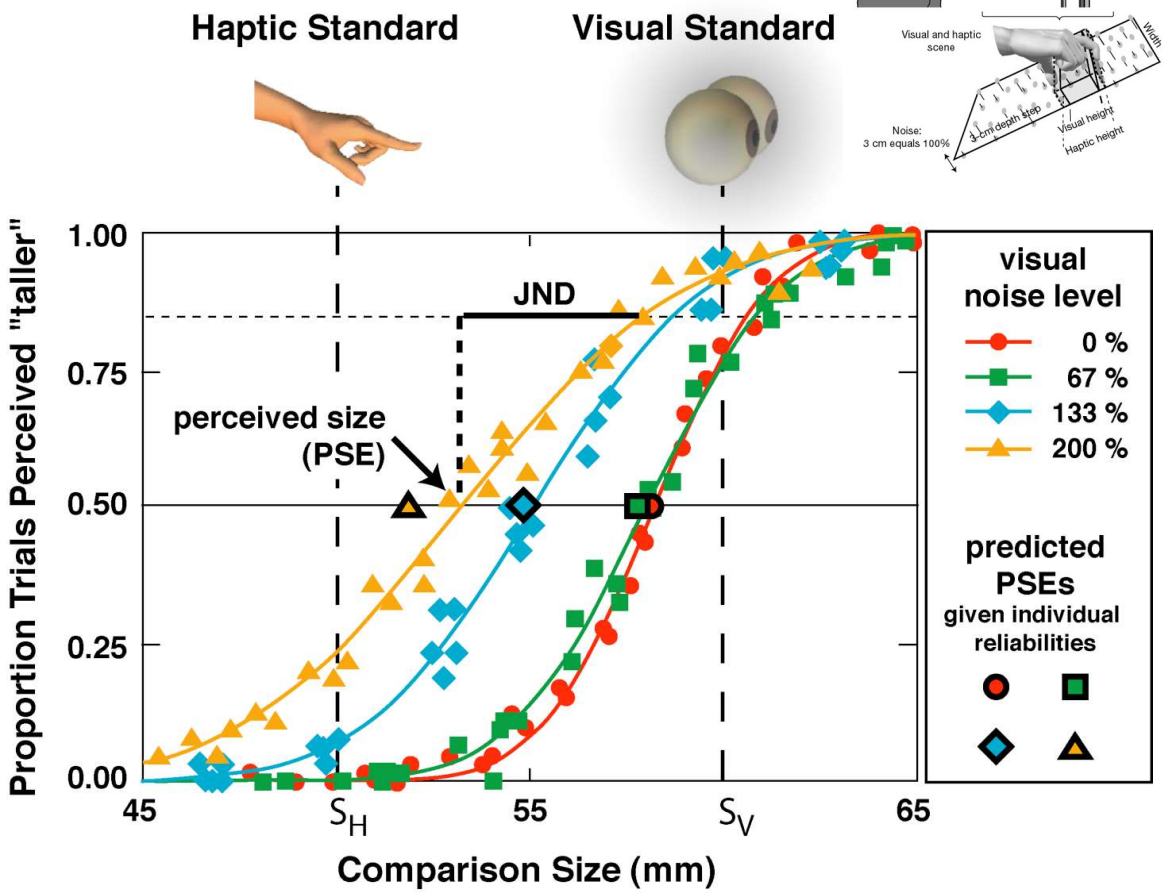


L'intégration Bayésienne de plusieurs indices sensoriels

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415(6870), 429-433.*



Résultats:
-la taille perçue (PSE, *point of subjective equality)* se déplace en direction de celle suggérée par la vision
- en proportion directe de la fiabilité des indices visuels
- avec un niveau de bruit (JND = *just noticeable difference*) qui décroit lorsque la fiabilité augmente



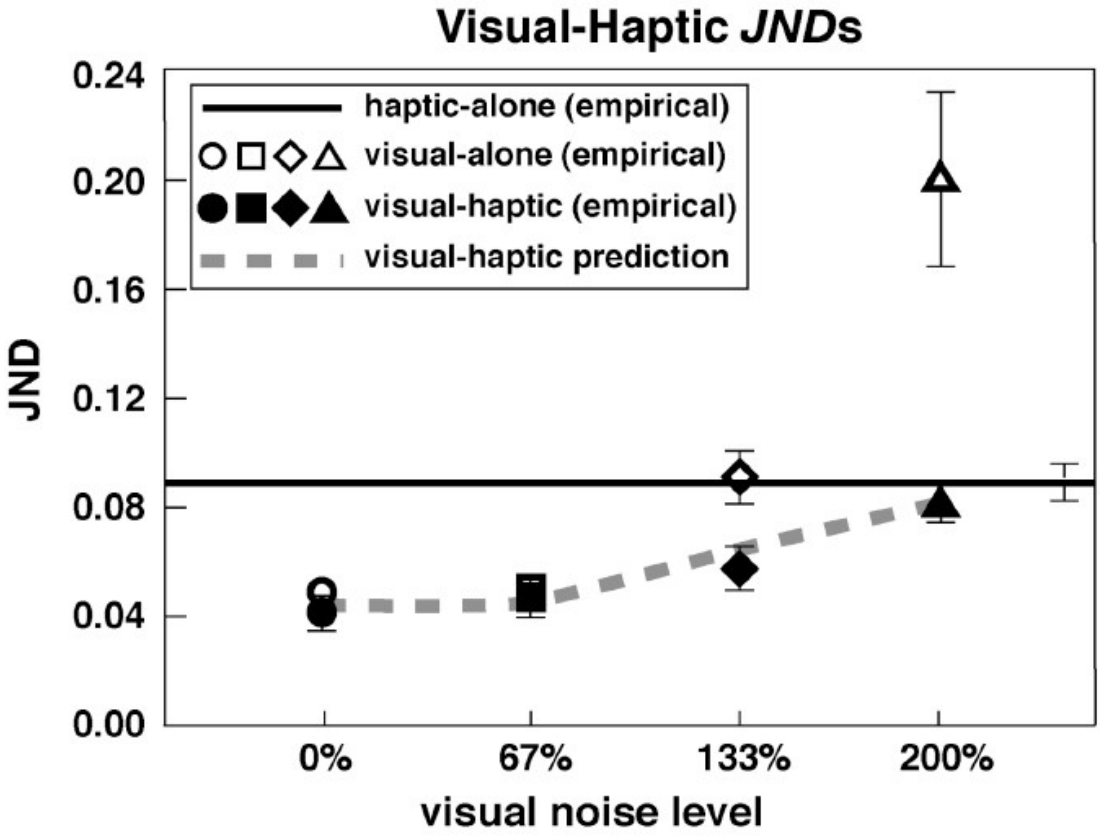
L'intégration Bayésienne de plusieurs indices sensoriels

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415(6870), 429-433.*

La précision de la réponse (JND = *just noticeable difference*) est quantitativement conforme aux prédictions de la théorie Bayésienne

Conclusion:

Le système perceptif intègre les sensations issues de deux modalités sensorielles selon les lois de l'inférence Bayésienne.



L'apprentissage: un autre domaine d'application de la théorie Bayésienne

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science, 331(6022), 1279-1285.*

« Comment nos esprits parviennent-ils à inférer autant à partir de si peu? » (Tenenbaum, *Science*, 2011).

C'est le « scandale de l'induction » discuté par Russell, mais aussi Platon, Aristote, Kant, Peirce…

Les enfants et les adultes réalisent quotidiennement des inférences très sophistiquées alors qu'il parait évident qu'ils n'ont pas assez de données.

Par exemple, tout le monde sait que « corrélation n'est pas causation » -- et pourtant les humains infèrent régulièrement des relations causales sur la base de quelques données qui ne suffiraient même pas à calculer un coefficient de corrélation!

Autre exemple: l'apprentissage du langage.

- Quine et le « gavagai ! »
- Chomsky et la « pauvreté du stimulus »

L’apprentissage: un autre domaine d’application de la théorie Bayésienne

Un exemple d’induction rapide (Tenenbaum, *Science*, 2011):

Les objects rouges sont des « tufa ».

L’induction Bayésienne dans l’apprentissage du langage:

- si les hypothèses sur le sens des mots sont des branches d’un arbre des catégories de sens envisageables
- alors la règle Bayésienne va automatiquement choisir la catégorie la plus petite, compatible avec les observations

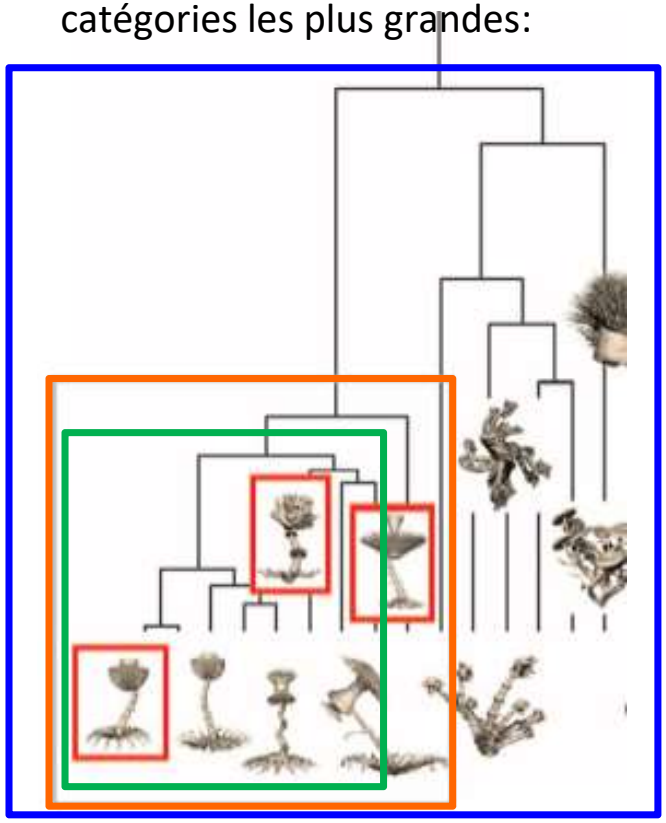
Supposons que toutes les hypothèses aient la même probabilité a priori.

$P(H|D_1,D_2,D_3)$ est proportionnelle au produit des vraisemblances $P(D_i|H)$ (en supposant que les observations sont conditionnellement indépendantes)

Les hypothèses **H** qui correspondent à des branches « trop petites » sont immédiatement éliminées : leur vraisemblance est nulle pour au moins l'un des mots: $P(D_i|\mathbf{H}) = 0$

Pour les autres catégories : $P(D_i|H) = 1/n$ où n est le nombre d'éléments de la catégorie

Le mécanisme de Bayes attribue automatiquement une vraisemblance plus faible aux catégories les plus grandes:



$P(D_i|\mathbf{H}) = 1/8$

$P(D_i|\mathbf{H}) = 1/14$

C'est l'une des versions du « rasoir d'Ockham ».

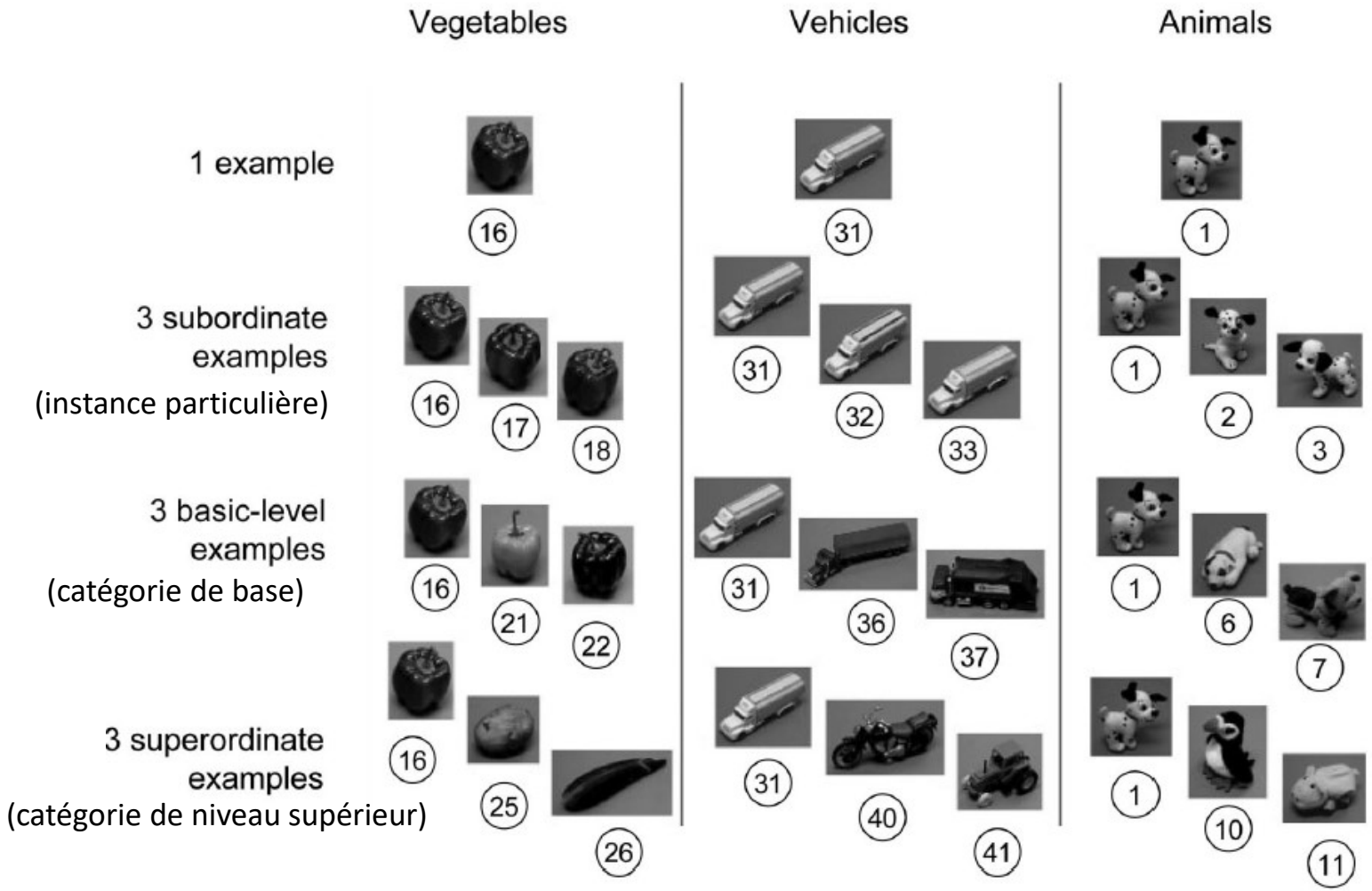
On attribue à Guillaume d'Ockham (1285-1347) un principe de raisonnement en réalité énoncé depuis l'Antiquité: « Une pluralité ne doit pas être posée sans nécessité »

« Les entités ne doivent pas être multipliées au delà du nécessaire »

Toutes choses égales par ailleurs, les explications les plus simples doivent être préférées aux plus complexes.

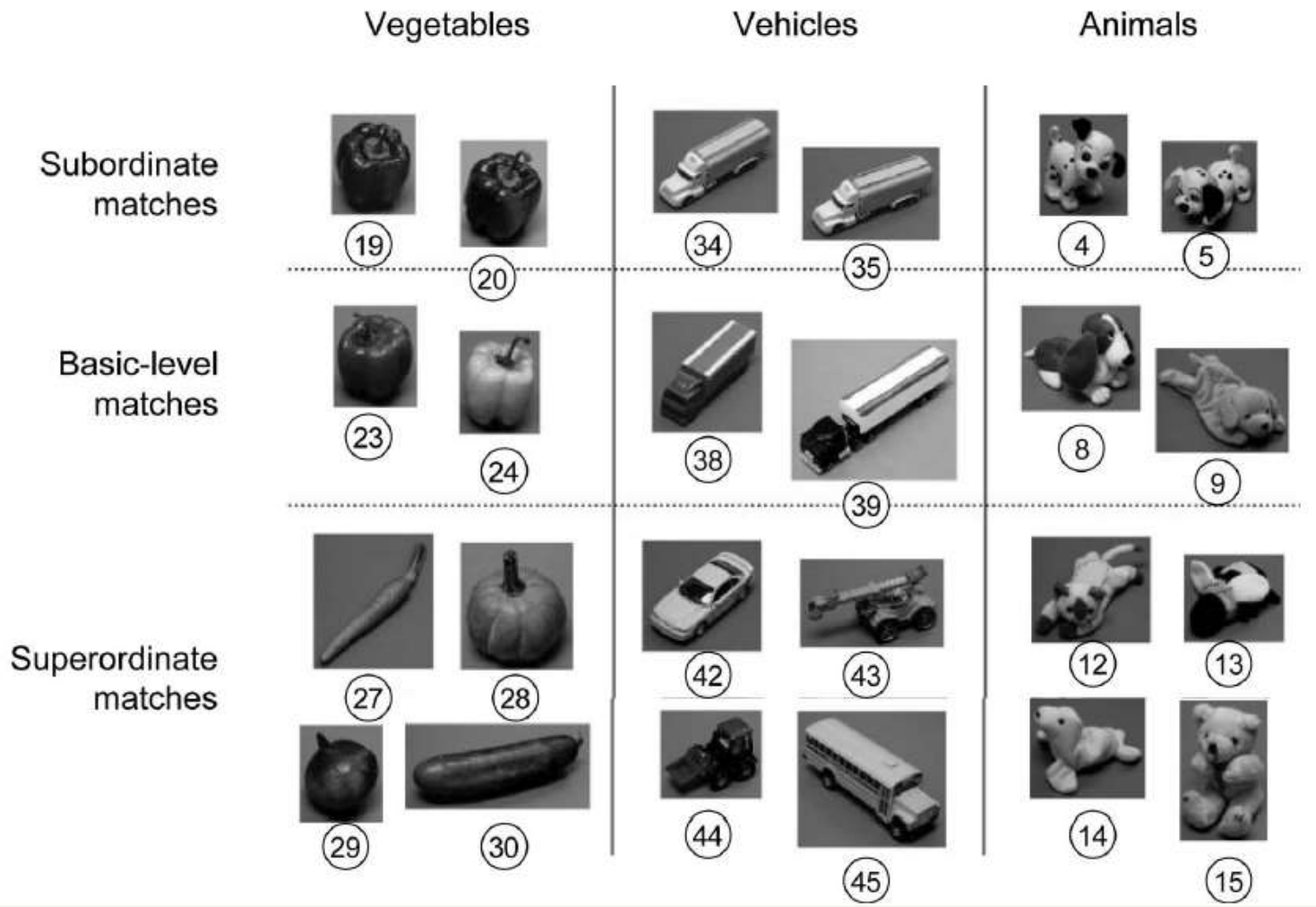
L’expérience de Xu et Tenenbaum (2007)

Les sujets sont exposés à divers objets, étiquetés par le même mot: « This is a fep ».
La diversité des exemplaires varie selon les conditions:



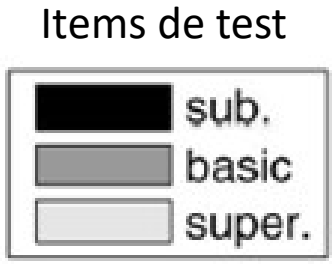
L’expérience de Xu et Tenenbaum (2007)

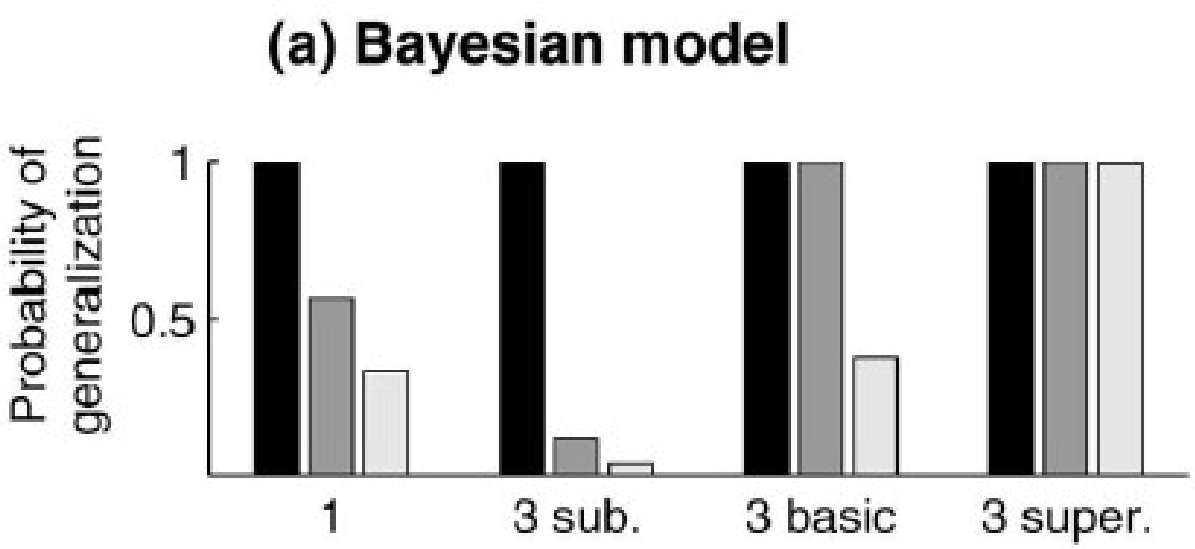
Puis on demande à la personne de généraliser: « Now tell me which of these items is a fep »



L'expérience de Xu et Tenenbaum (2007)

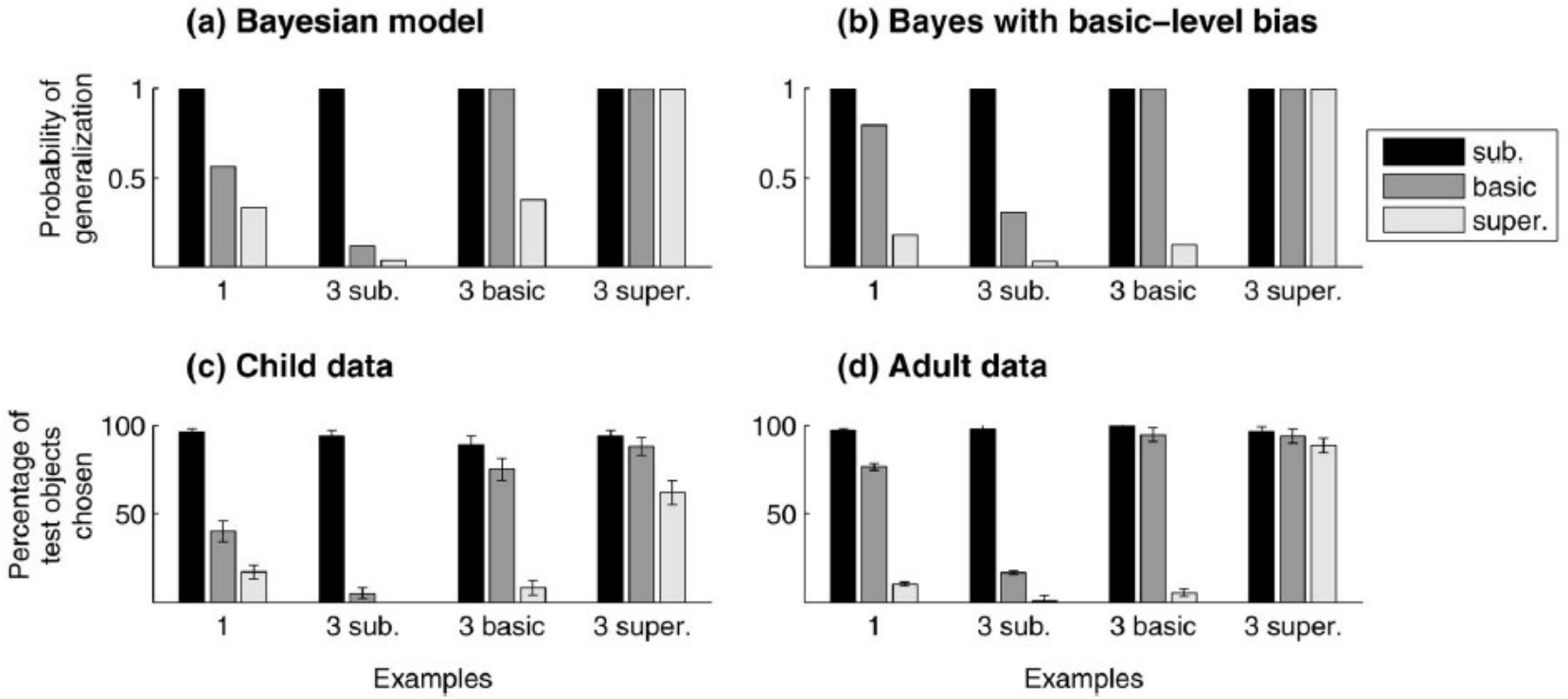
Prédictions du modèle bayésien:
(l'arbre des hypothèses est fondé sur une analyse en clusters des jugements de similarité entre les items).





L’expérience de Xu et Tenenbaum (2007)

Comparaison avec les résultats empiriques chez l’enfant de 3-4 ans et chez l’adulte.

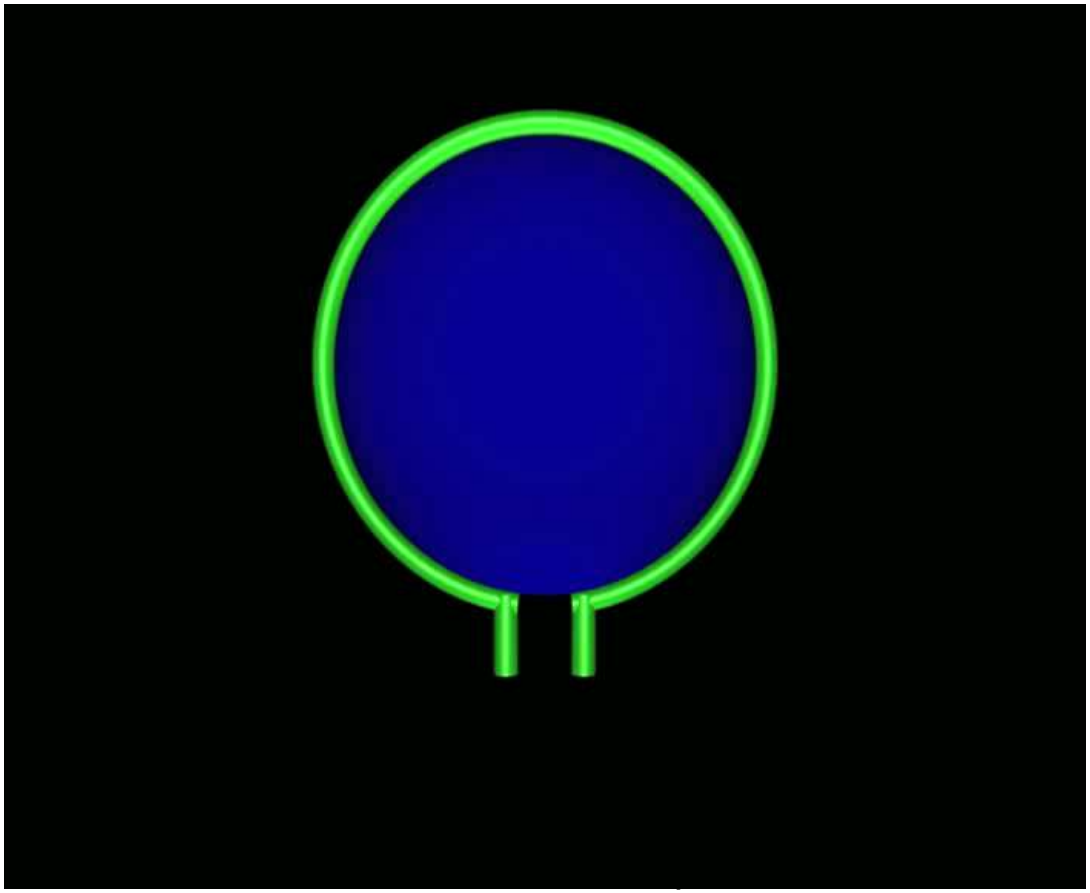


Deux visions du cerveau du bébé

Le bébé statisticien: les études de Teglas et Bonatti

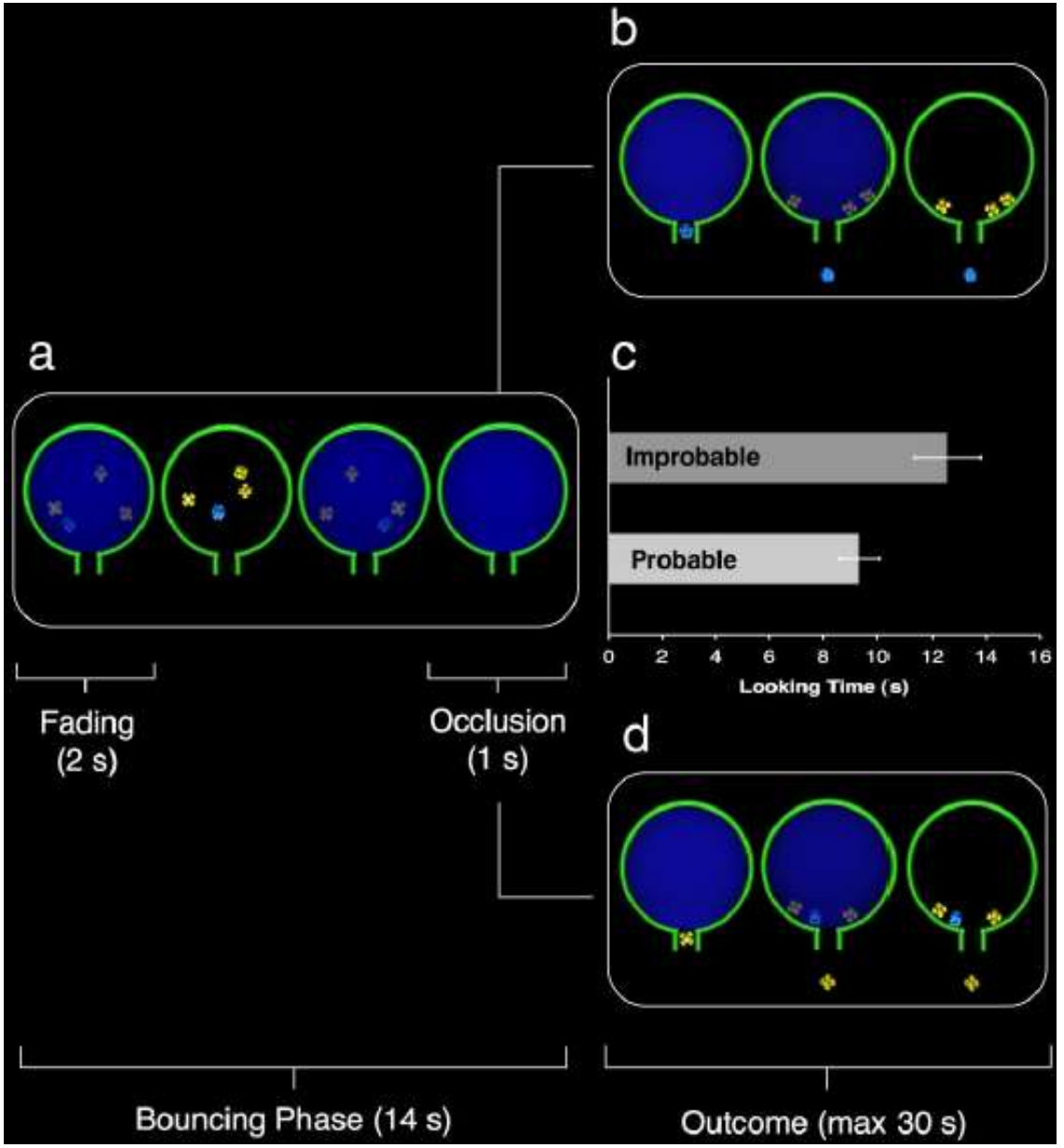
Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proc Natl Acad Sci U S A, 104(48), 19156-19159.*

Les bébés de 12 mois peuvent-ils anticiper la probabilité d'un événement futur?

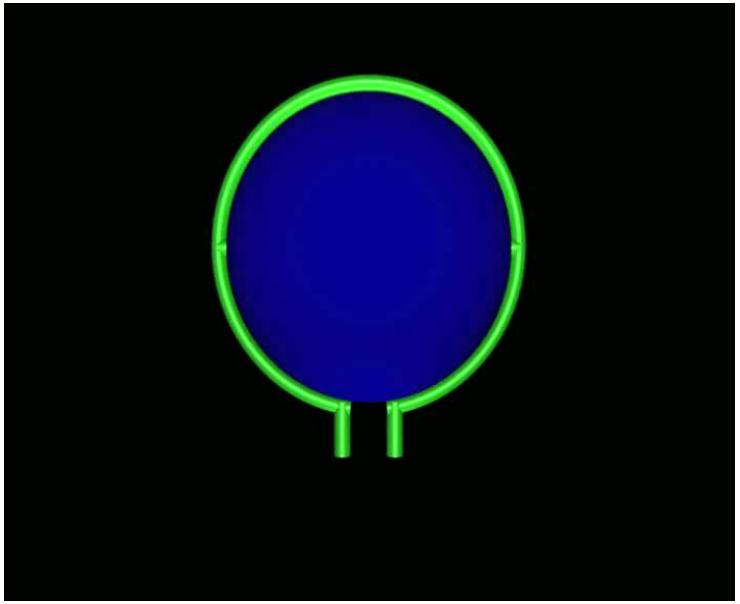


Résultats:

Les bébés de 12 mois regardent plus longtemps l'événement improbable que l'événement probable.

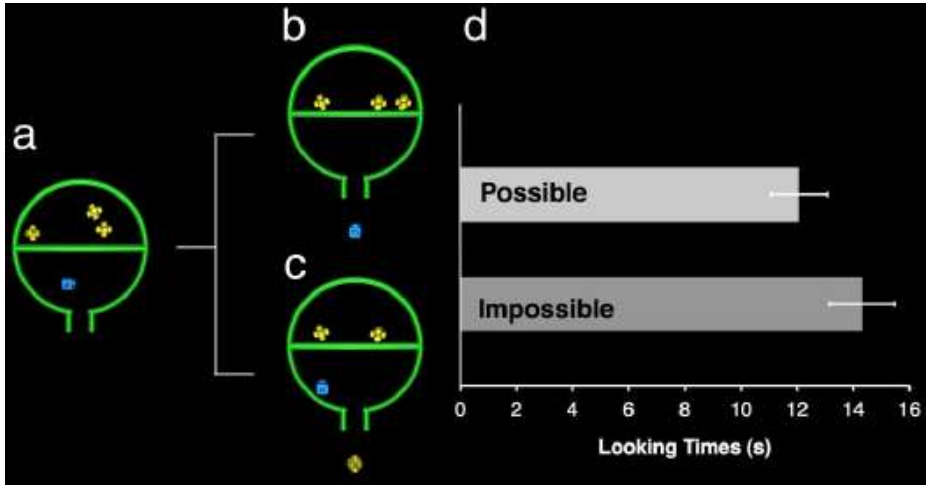


Ces résultats peuvent-ils s'expliquer par un simple biais perceptif?



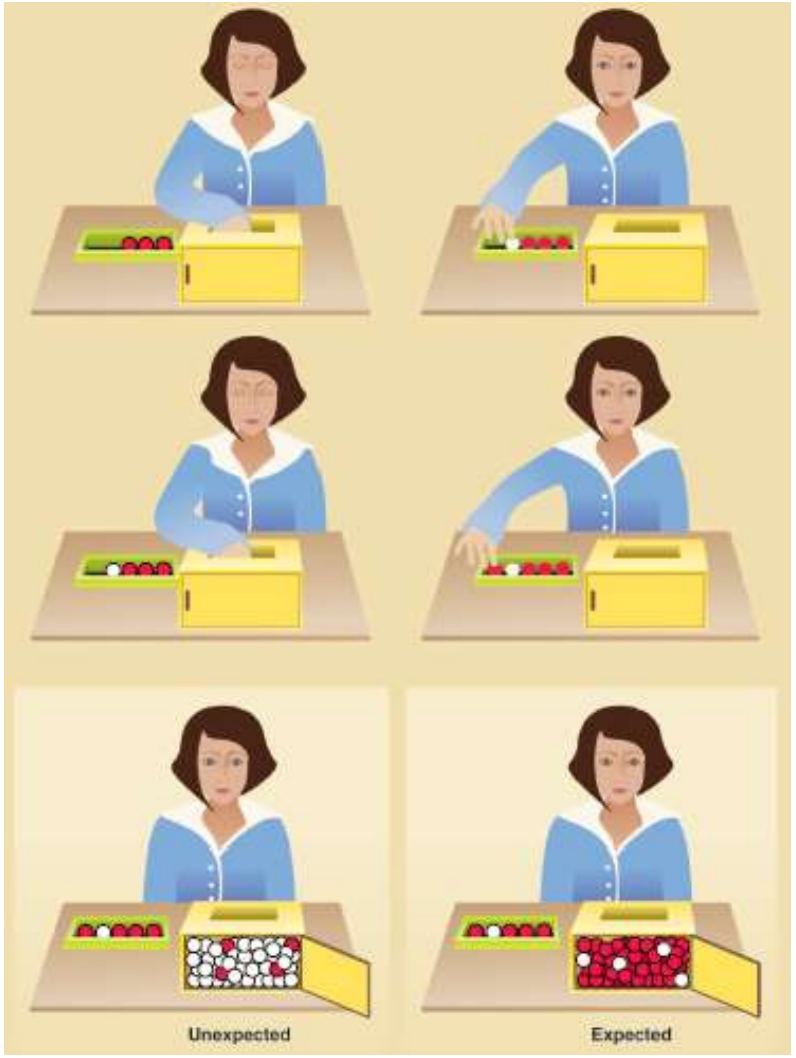
Un contrôle expérimental astucieux:
-le nombre d'objets est exactement le même qu'auparavant.
-une cloison rend impossible la sortie des objets les plus nombreux
-ainsi les résultats de l'expérience sont inversés: la seule issue possible est la sortie de l'objet unique

Résultats:
-la durée du regard s'inverse
-l'enfant regarde plus longtemps l'événement impossible, qui correspond à présent à la sortie d'un des objets les plus nombreux.



Une inférence bayésienne chez l'enfant: p(urne|échantillon) =?

Dans les expériences 1 et 2:
- On familiarise d'abord des enfants de 8 mois avec des balles rouges et blanches, et avec des boites qui contiennent, soit une majorité de balles blanches, soit une majorité de balles rouges.
-Puis, à chaque essai, l'expérimentateur montre une urne à l'enfant, mais son contenu n'est pas visible.
- Les yeux fermés, l'expérimentateur retire, une à une, cinq balles (4 d'une couleur, une de l'autre couleur).
- le contenu de l'urne est révélé
Résultat: l'enfant regarde plus longtemps l'urne dont le contenu ne correspond pas à l'échantillon observé.



Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proc Natl Acad Sci U S A, 105(13), 5012-5015.*

L'utilisation des statistiques pour détecter et apprendre des mots

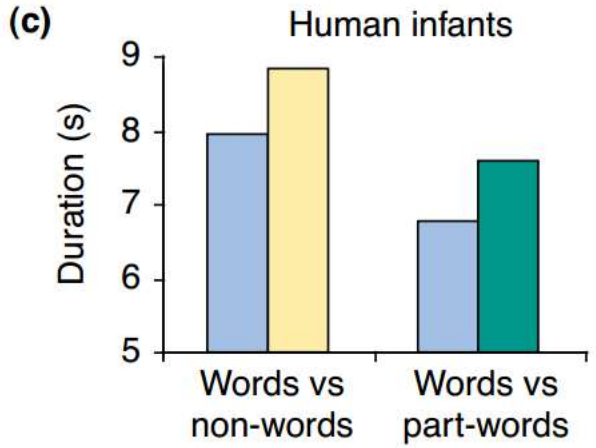
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–8.
Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53-64.
Review: Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, *5*(12), 539–546.

tokibugikobagopilatipolutokibu
gopilatipolutokibugikobagopila
gikobatokibugopilatipolugikoba
tipolugikobatipolugopilatipolu
tokibugopilatipolutokibugopila
tipolutokibugopilagikobatipolu
tokibugopilagikobatipolugikoba
tipolugikobatipolutokibugikoba
gopilatipolugikobatokibugopila

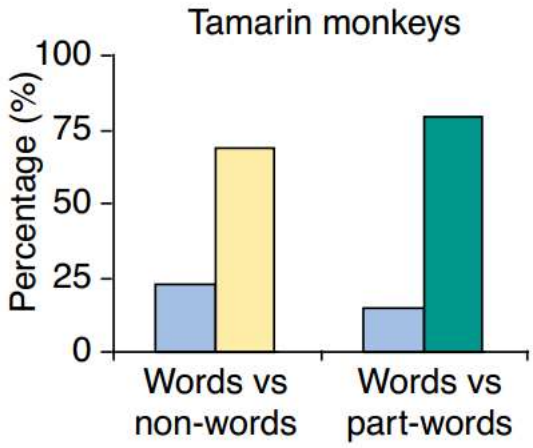
Les enfants de 8 mois font la différence entre « vrais » mots et les deux types de contrôle.
Les singes tamarins également !

Exposition à une séquence de syllabes formées d'un concaténation de « mots » de 3 syllabes.
Ensuite on fait écouter des chaînes de trois syllabes qui sont
-soit des « mots » de la séquence initiale: **tokibu**
-soit des non-mots: **gokiba**
-soit des « parties de mots » : la fin d'un mot et le début du suivant, par ex. **bugiko**

Durée d'orientation



Probabilité de s'orienter



Conclusion: Les idées essentielles de la théorie bayésienne

L'**inférence bayésienne** est une théorie mathématique simple: c'est l'extension de la logique classique au **raisonnement plausible** en présence d'incertitudes.

Nos **décisions** reflètent un calcul bayésien des probabilités (combiné avec une estimation des conséquences de nos choix) : p(réponse|stimulus)

L'inférence bayésienne rend bien compte des processus de **perception**: étant donné des entrées ambigües, notre cerveau en reconstruit l'interprétation la plus probable: p(objet|entrées sensorielles)

La théorie bayésienne explique bon nombre d'illusions visuelles (rôle de l'*a priori*) et également la manière dont nous fusionnons des indices multiples.

Le raisonnement bayésien fournit un **puissant algorithme d'apprentissage** de régularités statistiques.

Le **bébé** semble dôté, dès la naissance, de compétences pour le **raisonnement probabiliste**.

L'inférence peut être **hiérarchique** et donner rapidement accès à des connaissances abstraites (exemple de l'apprentissage des mots).

L'architecture du cortex pourrait avoir évolué pour réaliser, à très grande vitesse et de façon massivement parallèle, des inférences bayésiennes.