Esprits Numeriques

Séance 3

Jonathan Simon

Apercu

- 1) Turing
- 2) Qu'est-ce que c'est que de mettre en œuvre un calcul?
- 3) Putnam: L'argument de la réalisabilité multiple
- 4) Chalmers: L'argument de l'invariance organisationnelle (qualia absents, évanescents ou dansants)

Turing

Alan Turing

Turing

- Turing est surtout connu pour ses travaux sur la théorie de la calculabilité : il a identifié le problème de l'arrêt (die Entscheidungsproblem), a fourni des définitions concrètes de nombreux termes centraux de la théorie du calcul (comme celui d'un ordinateur universel, alias une machine de Turing). Il a également travaillé à la décodage de la machine Enigma (le système de cryptographie nazi) pendant la Deuxième Guerre mondiale.
- En parallèle, il a rédigé ce document, fixant l'agenda de la recherche sur l'IA pour les 50 prochaines années

Machines Informatique et Intelligence

- 1) Les machines peuvent-elles penser ?
- propose Turing : il est trop difficile de répondre à cette question. Remplacez-la par :
- 2) Une machine peut-elle être performante (tromper quelqu'un 30 % du temps) au jeu d'imitation ?

Machines Informatique et Intelligence

- Le jeu de l'imitation :

- Un examinateur, deux joueurs. Un joueur est humain, l'autre est une machine. Les joueurs sont dans des salles séparées, seuls des messages écrits sont échangés. L'examinateur peut demander n'importe quoi à l'un ou l'autre des participants, le but étant de deviner qui est l'humain. L'objet du jeu est de tromper l'examinateur en lui faisant croire que l'autre joueur est la machine.

Machines Informatique et Intelligence

- Pourquoi propose-t-il ce remplacement ?
- 1) Il évite un débat philosophique insolubles, le remplace par une question qui reste assez substantielle et difficile à répondre
- 2) Comportementalisme / Positivisme ?
- 3) Théorie cartésienne de la primauté de la preuve de la capacité linguistique (rappel de Descartes vs. Romanes et Huxley)

Machines Informatique et Intelligence

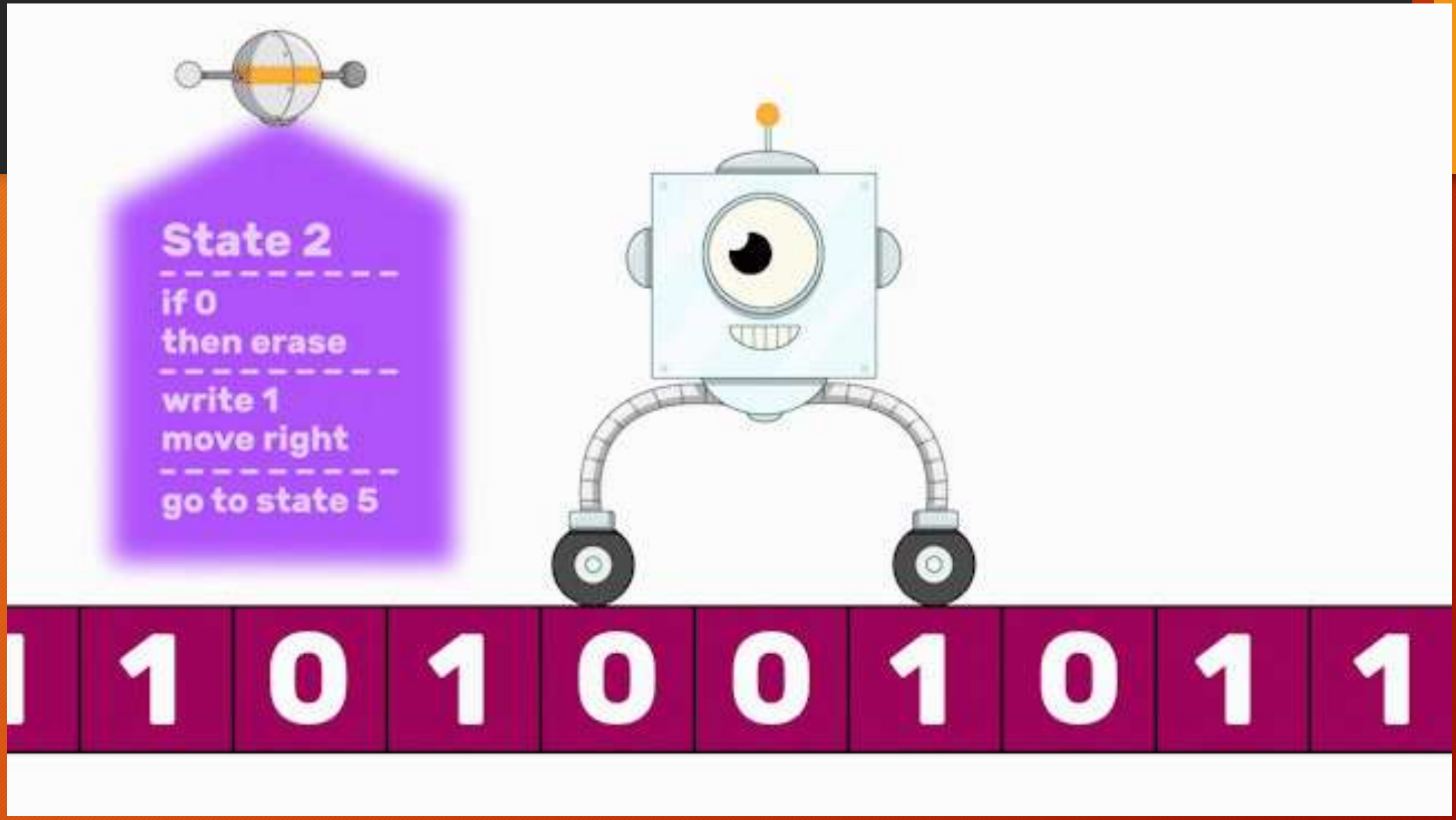
- Le reste du document :
- 1) la formulation de ce que doit être un ordinateur numérique (une machine de Turing) et de ce que doit être l'exécution d'un programme informatique (une machine à états discrets).

- 2) Ensuite, les réponses aux objections prévues pour affirmer qu'un ordinateur numérique sera capable de passer le test 50 ans plus tard (en 2001)

- 3) Enfin une esquisse positive d'une voie à suivre : ici, quelques réflexions classiques sur l'IA. Notez cependant qu'il ne les utilise pas pour répondre à des objections...

Les machines de Turing

- Le vrai génie de tout cela est qu'il est possible de démontrer que pratiquement toute opération formelle (qui intuitivement est quelque chose qui peut être fait sans nécessiter de perspicacité ou d'imagination, quelque chose que "n'importe quel ordinateur pourrait faire") peut être encodée arithmétiquement de la manière appropriée


State 2
if 0
then erase
write 1
move right
go to state 5
1 0 1 0 0 1 0 1 1

Le fonctionnalisme machinique

- Une machine de Turing : une longueur de bande / papier (aussi longue que nécessaire) divisée en cellules.
- Chaque cellule porte un numéro écrit dessus.
- Un lecteur/enregistreur fonctionne sur une cellule à la fois.
- Il lit le numéro, puis le modifie selon son *état*, et selon un ensemble de règles, puis se déplace vers une autre cellule (et modifie son état) selon un ensemble de règles.

Le fonctionnalisme machinique

- Il est essentiel que les règles indiquent au scanner ce qu'il doit écrire étant donné ce qu'il lit *et dans quel état il se trouve*. Les règles disent aussi *dans quel état il doit aller*.

Le fonctionnalisme machinique

La liste des règles est appelée *une table de machine*. Voici un exemple de table de machine pour une machine à coke (prochaine page). Notez que la «signification» des états est définie implicitement (les étiquettes ne sont ici qu'à des fins heuristiques)

Example: Machine à Coke (Coke coute 10c)

TABLE 1

	S_1	S_2
Input : pièce de 5 cents	N'émettre aucun output Passer en S_2	Emettre un coca-cola Passer en S_1
Input : pièce de 10 cents	Emettre un coca-cola Rester en S_1	Emettre un coca-cola et une pièce de 5 cents Passer en S_1

Example: Machine à Coke (coke coute 15c)

Input	État	Prochain État	Output
5c	Désire 15c	Désire 10c	rien
5c	Désire 10c	Désire 5c	rien
5c	Désire 5c	Désire 15c	Coke!
10c	Désire 15c	Désire 5c	rien
10c	Désire 10c	Désire 15c	Coke!
10d	Désire 5c	Désire 10c	Coke!

Example: Machine à Coke (coke coute 15c)

Input	État	Prochain État	Output
5c	S_0	S_1	rien
5c	S_1	S_2	rien
5c	S_2	S_0	Coke!
10c	S_0	S_2	rien
10c	S_1	S_0	Coke!
10d	S_2	S_1	Coke!

Le fonctionnalisme machinique

Observations :

1) Seules les entrées et les sorties sont définies explicitement, les états internes sont définis implicitement

2) Holisme : les états ne sont définis qu'implicitement en référence les uns aux autres. Si vous modifiez l'un des états, vous modifiez chacun d'eux

Turing: Machines Informatique et Intelligence

Réussir le test

- Quel est exactement son argument selon lequel une telle machine pourrait passer le test ? Il n'a pas vraiment d'argument décisif : il examine plutôt les objections et y répond.

- À noter ici : en particulier, il ne fonde pas explicitement tout sur l'idée que toute pensée humaine est une manipulation formelle de symboles : son objectif principal est de faire valoir qu'un ordinateur numérique pourrait réussir le test, et non qu'une "IA classique" pourrait. Les hypothèses qui conduisent à l'IA classique n'arrivent qu'à la fin

Objections

- Objection théologique
- La tête dans le sable
- Objection mathématique
- Argument de la conscience
- Divers handicaps
- L'objection de Lady Lovelace
- La continuité du système nerveux
- L'informalité du comportement
- Perception extra-sensorielle

Objections

- Objection théologique: Nous ne pouvons pas créer des âmes ! Turing : en effet. Mais ici, la question est seulement de savoir si nous pouvons créer des manoirs pour eux.
- La tête dans le sable: C'est effrayant ! Turing : pas d'objection
- Objection mathématique: Limitations formelles sur ce que les ordinateurs peuvent penser / prouver ! Turing : pourquoi de telles limitations ne peuvent-elles pas exister aussi pour les connaissances humaines ?

Objections

- Argument de la conscience: Une machine ne peut jamais vraiment être consciente ? Turing : vous pourriez changer d'avis si vous en voyiez une passer le test. Nous pouvons également aborder la question de savoir si la machine peut passer le test sans adresser celui-ci
- Divers handicaps: Mais une machine ne sera jamais capable de... Turing : soit non pertinent (manger de la glace), soit suppliant (réussir ce test), soit faux (faire des erreurs)
- L'objection de Lady Lovelace: Les machines ne peuvent pas vraiment créer / engendrer quoi que ce soit. Turing : dans un sens pertinent, il n'est pas évident que nous non plus.

Objections

- La continuité du système nerveux: Le système nerveux est continu et non discret. Turing : l'examinateur en jeu ne pourrait pas en tirer profit.

- (q : oui, mais cela pourrait-il avoir une importance pour que la machine soit suffisamment performante ?)

Objections

- L'informalité du comportement: Le comportement humain ne peut pas être décrit par un système de règles !

- Turing : Peut-être qu'elles peuvent être son juste difficile. Notez combien il serait difficile de déduire les règles qui régissent un ordinateur réel simplement en observant son comportement...
- Perception extra-sensorielle: ?

Apprentissage machine

- Turing conclut son article par quelques réflexions sur la façon d'amener les machines à apprendre à bien jouer (plutôt que de les programmer explicitement).

- Ses idées ici anticipent certaines idées dans l'apprentissage du renforcement, mais vont aussi clairement dans le sens de l'IA classique. Mais il ne s'appuie pas sur ces idées dans la première partie de l'article pour défendre ses hypothèses clés...

Compréhension / consolidation

- Est-il logique d'utiliser un modèle comme celui-ci pour tester la conscience plutôt que l'intelligence ?
- Ce test soutient-il vraiment le computationnisme, même s'il est en fin de compte comportemental ?
- Comment faire la différence entre un ordinateur qui fonctionne mal et quelque chose qui n'est pas du tout un ordinateur ?

Qu'est-ce que c'est que de mettre en œuvre un calcul?

Chalmers, Maudlin, Klein, Bostrom:

Bonnes et mauvaises abstractions

Bonnes et mauvaises abstractions

- Si les machines de Turing (en tant que descriptions abstraites ou mathématiques) spécifient des algorithmes, comment dire qu'un système donné «est» (ou, implémente / met en ouevre) une machine de Turing ?

Bonnes et mauvaises abstractions

- **Premier problème** : les descriptions des machines de Turing sont trop abstraites - même un rocher pourrait être «*interprété*» comme une machine de Turing
- **Deuxième problème** : les descriptions des machines de Turing sont trop exigeantes : même les ordinateurs ont des problems, font des erreurs, etc. Cela signifie-t-il que ce ne sont pas vraiment des ordinateurs ?

Bonnes et mauvaises abstractions

- Problème: les descriptions des machines de Turing sont trop abstraites - même un rocher pourrait être «interprété» comme une machine de Turing

Example: Machine à Coke (coke coute 15c)

Input	État	Prochain État	Output
5c	Désire 15c	Désire 10c	rien
5c	Désire 10c	Désire 5c	rien
5c	Désire 5c	Désire 15c	Coke!
10c	Désire 15c	Désire 5c	rien
10c	Désire 10c	Désire 15c	Coke!
10d	Désire 5c	Désire 10c	Coke!

Example: Machine à Coke (coke coute 15c)

Input	État	Prochain État	Output
5c	S_0	S_1	rien
5c	S_1	S_2	rien
5c	S_2	S_0	Coke!
10c	S_0	S_2	rien
10c	S_1	S_0	Coke!
10d	S_2	S_1	Coke!

Example: Machine à Coke (coke coute 15c)

Input	État	Prochain État	Output
I_1	S_0	S_1	O_1
I_1	S_1	S_2	O_1
I_1	S_2	S_0	O_2
I_2	S_0	S_2	O_1
I_2	S_1	S_0	O_2
I_2	S_2	S_1	O_2

Bonnes et mauvaises abstractions

- Problème: les descriptions des machines de Turing sont trop abstraites - même un rocher pourrait être «interprété» comme une machine de Turing

Bonnes et mauvaises abstractions

- Problème:

Étiquette l'environnement initial de la roche comme I1, sa configuration initiale S1, sa configuration un moment plus tard S2, l'environnement comme O1 : alors il correspond à une ligne sur ce tableau !

Bonnes et mauvaises abstractions

- Résolution:
- Nous avons besoin d'une sorte de contrainte contrefactuelle, ou d’isomorphism entre table de machine et mécanisme, c'est-à-dire que si le système avait été dans un état initial différent ou avait reçu une entrée différente, il aurait quand même satisfait à la description (de table de machine) en question...

Bonnes et mauvaises abstractions

- problèmes résiduels :
- 1) Est-ce trop fort ? Nous devons permettre que, par exemple, les ordinateurs puissent avoir des problèmes (tout en restant des ordinateurs). Quel degré de défaillance pouvons-nous autoriser ? À quel moment un système passe-t-il du statut d'ordinateur très glitchy ou peu fiable, à celui de ne plus être un ordinateur (Bostrom).

Bonnes et mauvaises abstractions

- problèmes résiduels :
- 2) Cause par omission : le fait de se fier à un critère contrefactuel peut impliquer que cela fait une différence, pour quelque chose qui ne se produit pas, si cela aurait pu se produire (l'alto silencieux ajoute à l'esthétique de la pièce, Klein).

Bonnes et mauvaises abstractions

- problèmes résiduels :
- 3) Ce n'est pas assez fort ? Nous pouvons décrire des systèmes qui respectent la condition contrefactuelle, mais qui intuitivement ne semblent toujours pas mettre en ouvre l'algo (Maudlin, Klein).

Bonnes et mauvaises abstractions

- problèmes résiduels :
- 3) Ce n'est pas assez fort ?
- Simon vs Theodore vs Alvin:

Bonnes et mauvaises abstractions

- Simon est une machine de Turing qui, étant donné un nombre, produit ses facteurs. Pour le nombre 20, elle produit (1,2,4,5). Étant donné le nombre 21, elle produit (1,3,7)...

Bonnes et mauvaises abstractions

- Théodore ne peut résoudre qu'un seul problème (par exemple, trouver les facteurs de 20 : 1,2,4,5). Quel que soit le nombre que tu lui donnes, il exécute la recette pour factoriser 20 et obtient (1,2,4,5).

Bonnes et mauvaises abstractions

- Alvin est composé de Théodore et de Simon : par défaut, il exécute le programme le plus simple de Théodore. Mais il possède un interrupteur qui vérifie si l'entrée est 20 ou non. Si l'entrée n'est pas 20, l'interrupteur désactive le mécanisme de Théodore et active celui de Simon.

- Alvin est contrefactuellement sensible.... mais si on lui donne l'entrée 20, il fait la même chose que Simon.

Putnam

Putnam

- Maintenant : arguments selon lesquels les robots / IA peuvent être conscients

- À la lumière de ce que nous venons de voir, note que cette affirmation est vague. On pourrait peut-être décrire correctement un humain comme mettant en œuvre un algorithme "IA"... mais les autres choses mettant en œuvre cet algorithme ne sont pas conscientes.

Putnam

- Notre question est la suivante : qu'est-ce qui est ***nécessaire et suffisant*** pour la conscience (par exemple pour ressentir de la douleur) ?

- … n'oublie pas que nous pouvons prendre les terme «necessaire» et «suffisant» dans le sens de «necessaire et suffisant compte tenu des lois psychophysiques » (l'approche dualiste) ou «necessaire et suffisant compte tenu des lois métaphysiques » (l'approche matérialiste)…

Putnam

- Putnam a été l'un des premiers à proposer une conception explicitement fonctionnaliste des états conscients (en réponse aux alternatives behavioristes et physicalistes)

- Contre les béhavioristes : son argument le plus célèbre (dans un autre article) est l'argument des Superspartans : des agents qui ne manifestent pas de comportement douloureux et qui, en fait, s'entraînent à ne plus avoir de dispositions pour le comportement douloureux.
  - (une version de l'argument selon lequel les conséquences comportementales des états mentaux, y compris les états phénoménaux comme la douleur, sont interconnectées de manière holistique).

Putnam

- Contre les physicalistes :

- Ici, les adversaires de Putnam sont ceux qui insistent sur le fait que le substrat compte, que ce n'est qu'un type spécifique d'état neuronal-physique-chimique qui peut provoquer l'expérience de la douleur....

Putnam

- «Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that any organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. At the same time, it must not be a possible (physically possible) state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain that it will also be a state of the brain of any extra-terrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may be pain.

Putnam

- It is not altogether impossible that such a state will be found. Even though octopus and mammal are examples of parallel (rather than sequential) evolution, for example, virtually identical structures (physically speaking) have evolved in the eye of the octopus and in the eye of the mammal, notwithstanding the fact that this organ has evolved from different kinds of cells in the two cases. Thus it is at least possible that parallel evolution, all over the universe, might always lead to one and the same physical "correlate" of pain. But this is certainly an ambitious hypothesis.

Putnam

- Finally, the hypothesis becomes still more ambitious when we realize that the brain-state theorist is not just saying that pain is a brain state; he is, of course, concerned to maintain that every psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say "hungry"), but whose physical-chemical "correlate" is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly probable that we can do this. Granted, in such a case the brain-state theorist can save himself by ad hoc assumptions (e.g., defining the disjunction of two states to be a single "physical-chemical state"), but this does not have to be taken seriously. »

Putnam

- 1) Nous pouvons trouver au moins un prédicat psychologique qui peut clairement être appliqué à la fois à un mammifère et à une pieuvre (disons «faim»), mais dont le «corrélat» physico-chimique est différent dans les deux cas.
- 2) Si 1), la théorie de l'état-cerveau est fausse
- 3) Par conséquent, la théorie de l'état-cerveau est fausse

Putnam

- Comment justifier 1): ??

- 1) Nous pouvons trouver au moins un prédicat psychologique qui peut clairement être appliqué à la fois à un mammifère et à une pieuvre (disons «faim»), mais dont le «corrélat» physico-chimique est différent dans les deux cas.

Putnam

- Comment justifier 1): ??
- En particulier, comment savons-nous que le même prédicat s'applique à la pieuvre?

Putnam

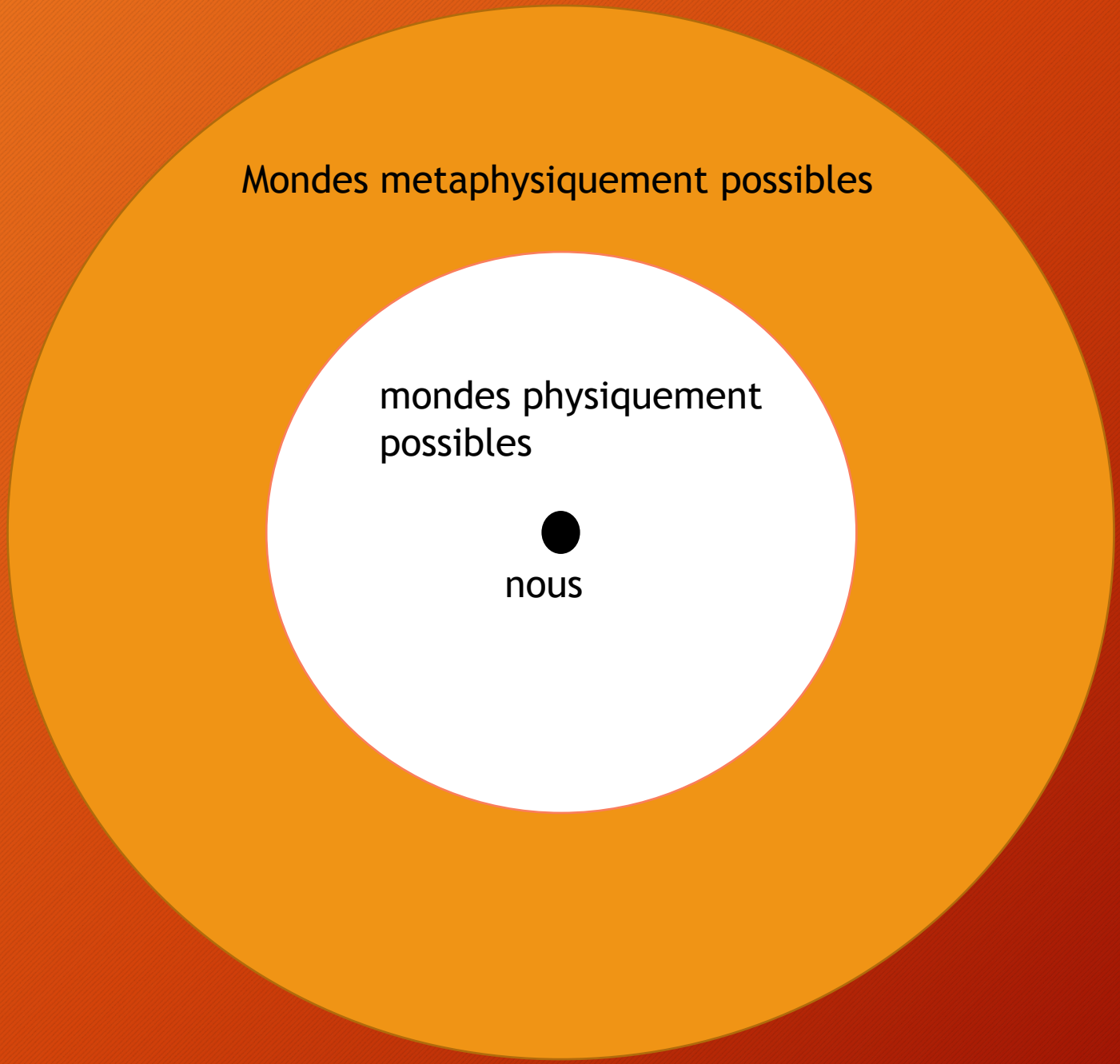
- Putnam: «we identify organisms as in pain, or hungry, or angry, or in heat, etc., on the basis oftheir behavior…»

- (Les similitudes comportementales sont la preuve de similitudes mentales. (Mais le théoricien de l'état du cerveau ne peut-il pas l'accuser de contourner la question ici ?)

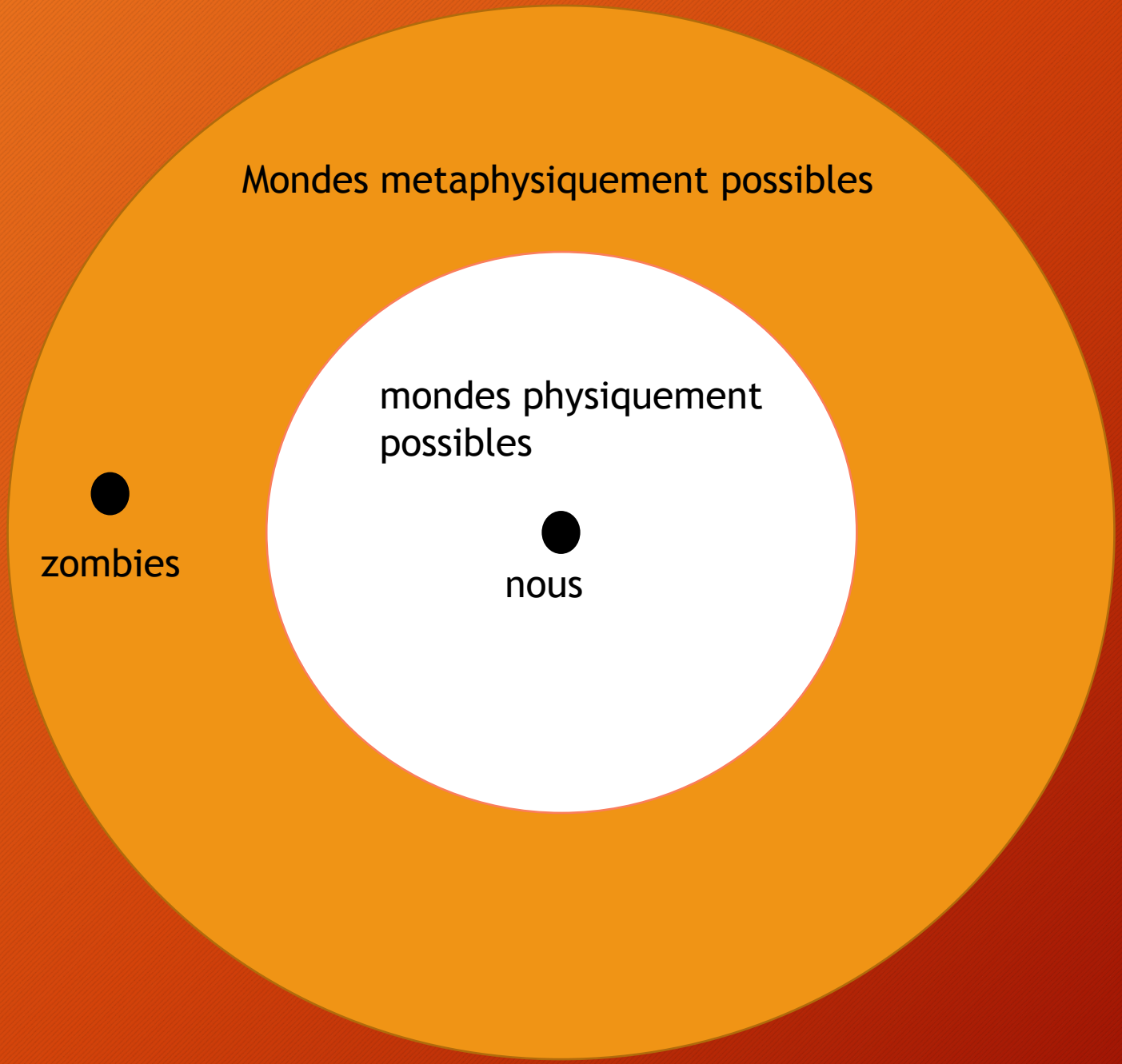
Chalmers

Chalmers

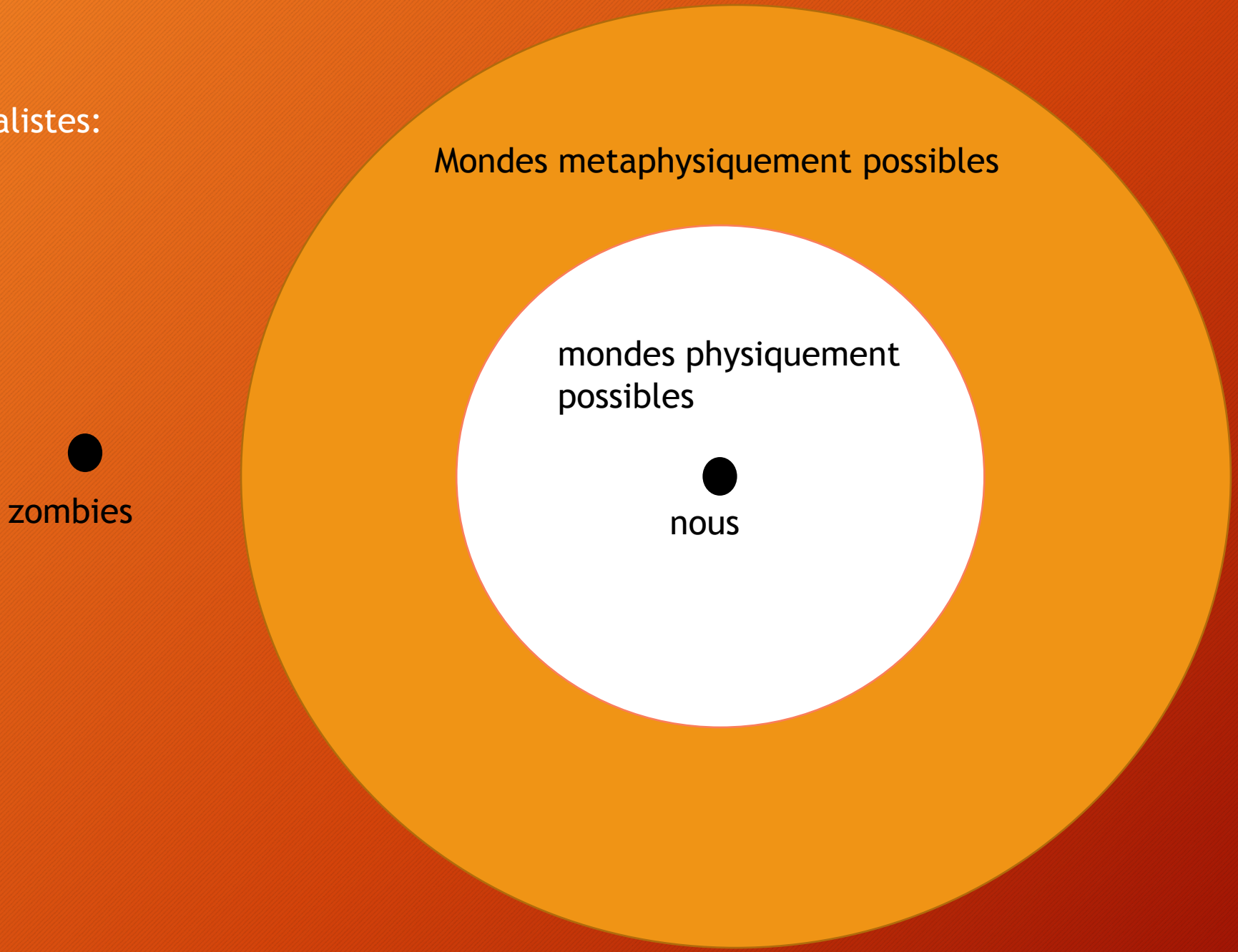
- 1) Chalmers est un dualiste des propriétés : il soutient que des propriétés comme «avoir mal» sont irréductibles à des propriétés physiques ou fonctionnelles - et en particulier qu'il existe des mondes possibles où il y a de parfaits doubles physiques et fonctionnels de nous, qui n'instancient pas ces propriétés (ne sont pas conscients) - zombies.... !


Mondes metaphysiquement possibles
mondes physiquement possibles
nous

Dualistes:


Mondes metaphysiquement possibles
mondes physiquement possibles
zombies
nous

Matérialistes:



Chalmers

- En physique (et dans d'autres sciences), on cherche des lois. Ces lois énoncent quelque chose de plus que la façon dont les choses se sont accidentellement produites dans les faits, mais quelque chose de moins que la façon dont les choses doivent être métaphysiques : elles énoncent la façon dont les choses doivent être dans les mondes naturellement possibles.

Chalmers

- Beaucoup considèrent que les vertus théoriques telles que l'élégance et la simplicité sont des guides vers la vérité : face à deux théories candidates qui expliquent les faits, nous devrions préférer celle qui est la plus simple

Chalmers

- Pour les dualistes, les principes qui relient les propriétés phénoménales (qualia) aux propriétés physiques vont être des lois (car il n'y a pas de connexions métaphysiquement nécessaires, ou du moins pas beaucoup).

Chalmers

- Les arguments de Chalmers visent à montrer que tout ensemble de lois (psychophysiques) qui nie l'invariance organisationnelle, sera moins simple que les lois que l'on peut avoir si on la préserve (où OI est la thèse selon laquelle la duplication fonctionnelle signifie la duplication phénoménale, dans les mondes naturellement possibles).

Chalmers

- Son argument (pour le fonctionnalisme) est intéressant en partie parce qu'il ne nécessite pas d'identité ou de réduction : au lieu de cela, il soutient que les lois psychophysiques (reliant les domaines distincts du physique et du mental) sont écrites au niveau des fonctions (computationnelles)

Chalmers

- Pour la conversation : si nous rejetons le dualisme, pouvons-nous encore nous appuyer sur son argument pour établir le fonctionnalisme ?

Chalmers

- (oui: en effet il *utilize* son dualisme pour établir la possibilité de cas de qualia absents ou inversés. (mais les théoriciens de l'identité cerveau-état devraient aussi l'accepter)

- (non: Tout cela est basé sur un appel à la simplicité des lois, mais si le réductionnisme est correct, nous ne cherchons pas des lois, mais des principes d'identité...)

Chalmers

- Les arguments:

- Des qualia évanescents contre des qualia absents

- Des qualia dansants contre des qualia inversés

Chalmers

- Qualia absents : doublons fonctionnels (articifiels) de nous qui n'ont pas d'expériences.

- Qualia inversés : doublons fonctionnels (artificiels) de nous qui ont des expériences différentes des nôtres.

Qualia évanescents

- 1) Il est possible qu'il existe un robot, fonctionnellement isomorphe à moi, Z, sans qualia
- 2) Si 1), alors il est possible qu'il y ait une série de cas, chacun avec un neurone remplacé par une puce en silicium, entre moi et Z
- 3) soit le passage de mon expérience à aucune expérience est soudain, sur une seule étape, soit il est progressif (qualia évanescents)

Qualia évanescents

- 3) soit le passage de mon expérience à aucune expérience est soudain, sur une seule étape, soit il est progressif (qualia évanescents)
  - 3a) si soudain: «Des discontinuités brutales dans les lois de la nature, contrairement à celles que l'on trouve partout ailleurs »
  - 3b) si progressif: alors le sujet se trompe sur (presque) tout ce qu'il vit, violant les principes naturels qui relient la conscience et la cognition (ce qui rend les lois qui les relient très compliquées).

Qualia évanescents

- 4) Dans tous les cas, les lois de la nature seraient très laides : plus laides que s'il n'y avait pas de qualia effervescents.
- 5) Dans tous les cas, les lois de la nature seraient très laides : plus laides que si les qualia effervescents n'existaient pas.
- 6) Les lois ne sont pas aussi laides
- 7) Par conséquent, il n’y a pas de qualia effervescents – preuve par contradiction que 1) et faux

Qualia dansants

- 1) Il est possible qu'il existe un robot, fonctionnellement isomorphe à moi, Z, avec qualia differents
- 2) Si 1), Il doit donc y avoir deux systems A et B qui diffèrent au maximum par un dixième de leur composition interne (neurone vs silicone), mais qui ont des expériences significativement différentes
- 3) Il est possible de prendre un circuit en silicone comme celui de B et de l'installer comme circuit de secours dans A (activé par un interrupteur).

Qualia dansants

- 4) Le basculement de cet interrupteur ferait danser les qualia de A, sans que A s'en aperçoive (même si nous supposons que c'est l'aspect auquel A prête attention)
- 5) Si 4), alors les lois reliant la conscience et la cognition sont très complexes et inélégantes
- 6) les lois reliant la conscience et la cognition ne sont pas complexes et inélégantes.
- 7) Donc,) Il n’est pas possible qu'il existe un robot, fonctionnellement isomorphe à moi, Z, avec qualia differents