NEURODEMOCRACY

SELF-ORGANIZATION OF THE EMBODIED MIND

A THESIS SUBMITTED
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

TA-LUN HUANG
THE UNIVERSITY OF SYDNEY
UNIT FOR HISTORY AND PHILOSOPHY OF SCIENCE

JUNE 2017

THE UNIVERSITY OF SYDNEY

Higher Degrees by Research Examiner's report on thesis

Report due date: 16/05/2017

Name of examiner: Professor Bryce Huebner

Examiner's institution: Georgetown University

Name of student: Ta Huang SID: 311057411

Faculty: Science Date of submission: 28/02/2017

Title of thesis: Neuro-Democracy: Self-Organization in the Embodied Mind.

Degree: PhD

A CONTENT AND PRESENTATION OF THESIS

As described in Clause 8 of Thesis and Examination of Higher Degrees by Research Policy 2015 , the thesis should:

a) be the student's own work, embodying the results of the work undertaken by the student during candidature;
b) form a substantially original contribution to the knowledge of the subject concerned;
c) afford evidence of originality by the discovery of new knowledge; and the exercise of independent critical ability;
d) form a cohesive and unified whole;
e) include a substantial amount of material that may be suitable for publication;
f) satisfactorily demonstrate that the student is able to identify, access, organise and communicate new and established knowledge;
g) be written in a standard generally acceptable to the discipline;
h) be written in English (except where permitted under the University of Sydney (Higher Degree by Research) Rule 2011)

B RECOMMENDATION

Please tick no more than one of the following recommendations:

After examination of the thesis and supporting material, I recommend that

(1) ***The student be awarded the degree without further conditions*** .... ☒

or (2) ***The student be awarded the degree subject to corrections of the thesis to the satisfaction of the University***.. ☐
Corrections include errors or omissions in the thesis, such as incorrect citations, omissions, or typographical errors, which must be corrected but which do not alter the conclusions of the thesis, Corrections may also include additions or deletion of material in the text, tables, figures, or appendices. Changes should, in the opinion of the examiner, not require an additional period of research and should not result in the conclusions of the thesis being significantly altered. Corrections do not require the thesis to be returned to examiners, but can be adjudicated by the Chair of Examination.

or (3) ***The student not be awarded the degree, but be permitted to resubmit a revised thesis for examination following a further period of study***. ..... …☐
The thesis in its current form does not merit award. There are errors and/or deficiencies that, in the opinion of the examiner, substantially affect the argument or the conclusions of the thesis. Changes may include but are not limited to the provision of extra data or material. The student demonstrates sufficient ability that, after an additional period of study, a thesis of the required standard may be achieved. Following revision and resubmission, the thesis is re-examined.

I feel a further period of research extending over ... . months would be necessary.

Please indicate whether you would be willing to re-examine a revised thesis if so invited.... Yes ☐ No ☐

Note*: This option is not available for a thesis that has already been revised and resubmitted for examination.*

or (4) ***The student not be awarded a doctoral degree but be awarded another degree for which they are eligible***…☐
The thesis is not considered satisfactory for the award of the degree for which it was submitted, but another degree for which the student is eligible may be awarded instead.

or (5) ***The student not be awarded the degree*** ..... …☐
The thesis does not merit award of the degree and does not demonstrate sufficient ability by the student for a resubmitted thesis to achieve this merit. For example, the hypothesis and methods may be fatally flawed, therefore rendering the conclusions completely invalid and not capable of being rectified by an additional period of study.

C RELEASE OF EXAMINER'S NAME AND COMMENTS TO THE STUDENT

THE UNIVERSITY OF SYDNEY

I agree that the University may release my name and report to the student during or after the examination or as required by legislation. [In exceptional circumstances, the University may withhold information about the examiner]

Examiner name Bryce Huebner .....13 April 2017 .....Date

Signature…  ……..

D GROUNDS FOR RECOMMENDATION

Please complete the following sections. In cases where examiners are not unanimous in their recommendation, this information will form the basis of the decision made by the University on the award or non-award of the degree. *Please note that the following boxes will expand according to the amount of text you type.*

Is the ***content and presentation*** of the thesis consistent with the description in Section A of this form?

Yes……☒ No……☐

Please state below the grounds on which you base your recommendation. You may also wish to provide suggestions for the next steps in research or improvements for publication that are not required for the award of the degree.

This is the strongest and most thoroughly argued thesis that I have ever read. The argument of the thesis hangs together nicely as a unified project, and the flow of the argument builds in a very interesting way. There were a few points where I thought that things could be argued in a different way, and there were a few points where I thought that parts of the argument could be dropped—making the main line of argument much clearer. But overall, the issues that I had with the project were cosmetic, and I think that they will be extremely easy to clean up as this thesis is prepared for submission as a book manuscript. I think that with a few changes, deletions, and additions, this project will make for a competitive submission to an excellent press (e.g., Oxford University Press or MIT Press, both of whom I referee for quite often). Finally, I hope that Linus Ta-Lun Huang makes these revisions quickly, and gets this book out, as this is a book that will be incredibly timely, and he is positioned to have a big effect on how the philosophy of cognitive science unfolds in the near future. So I offer the following suggestions as ways of preparing this project for publication.

- Overall, there is way too much signposting in the thesis. While it is helpful to keep tabs on what has happened, and what is going to happen next, too much time is devoted at each of the section transitions to making these kinds of things clear. This is not a problem in the context of a thesis. However, the flow of the argument needs to be streamlined a bit to make this a readable book. And I would suggest limiting these kinds of sign posts, and packing the overview of the book into the latter half of an introduction.

- At many points in the thesis, there are quick discussions of embodiment. And it is even suggested that embodiment is the right place to push for a theory of cognitive architecture. However, embodiment—as such—plays a very small role in the arguments that are developed in the thesis. To be honest, I thought that the discussion of embodiment was more of a distraction than anything, and the view that emerges is more concerned with linking perception and action than it is with things like the role of embodied states, or even the role of interoceptive processing in cognitive architecture. My strong recommendation is to drop the term embodiment, and to drop the parts of the argument that focus on embodiment, and instead make it clear that this is an action-oriented cognitive architecture, with close ties to theories of affordances. That will help to make more people appreciate the importance of the book; and it will make it less likely that people will get off board because they find claims about embodiment unappealing. The story in the thesis is consistent with a minimally-embodied perspective on cognitive science, and focusing on embodiment is likely to cause more problems than it solves. Put somewhat differently, the real issue in the book is *control*, and in the version that becomes a book it will be good to focus on that.

- Given the current prominence of predictive processing accounts of cognition, I think that it would be helpful to bring them into the project much earlier. As I see it, these approaches have attempted to solve the control problem, and they have done so by developing an account of cognitive architecture. Hence, the framing on p.2 comes off a bit strong, and it feels like this vibrant area of research is being ignored. It is discussed later in the project, but it would help to make it clear that this project is an attempt to make good on some of the issues regarding cognitive architecture that are starting to be discussed more commonly, rather than framing this as a completely new way of wading into this project.

- In a book, Chapter 2 could be compressed quite a bit. It's very much an introduction to issues of modularity, and a rapid fire overview of a bunch of architectural and epistemological details of different positions. But it doesn't push things ahead very much. It was good in a thesis, as it shows an understanding of the key issues here. But I think that it would seem like too much orienting in the context of a book.

- In the discussion of modularity and learning in Chapter 3, it seemed to me like there were two positions that would need to be discussed in the context of a book on this topic. The first is the sort of modularisation that Annette Karmiloff-Smith famously argued for. I think that a view like this allows for a lot more learning and flexibility than

the kinds of nativism that are discussed in this thesis. But I'm not quite sure how they fit with the kinds of arguments that are being developed here, and I would like to see a discussion of this in the book. Second, it would be helpful to think of some of the related ideas that have recently been developed by Cecilia Heyes, in her discussion of learning and development, and Stan Dehaene in his account of reading. Both of these kinds of positions allow for a great deal more complex 'modular' structure, and before the classical view is dismissed, these kinds of positions should be addressed.

- The discussion of massive modularity as it is developed by evolutionary psychologists was a bit quick, and in a fleshed out version of chapter 3, it would help to spell out some examples of how they think that central modules work, and where it is that we should expect them. Maybe a discussion of the well known stuff on deontic reasoning could work here; but I think that it would be far more useful to offer a discussion of their claims about racial bias, and why we shouldn't expect it to be an adaptation—along with their discussion of own-group bias, which we should expect as an adaptation.

- Also in chapter 3, it would really help to say a bit more about how flexible human cognition actually is, and what it means to be committed to the kind of cognitive flexibility that is discussed around p.48ff. I found the discussion of Poverty of Stimulus arguments to be really clear and really helpful, but one of the core claims that needs to be addressed in more detail is the extent to which some trajectories are supposed to be unlearnable. I gather that part of the reason that nativists end up being nativist is that they think that there are strong constraints on the kinds of capacities that we can develop. Languages, for example, don't develop in random ways—they all fall within a fairly narrow range of variation. So in making claims about cognitive flexibility, it needs to be shown that there *is* more flexibility than can be allowed by a nativist project. And this is going to require saying a bit more about the domains of flexibility that are at issue, and the kinds of constraints that would have to be imposed in a problematic way by nativists. (In parallel, one of the main claims that it is advanced by proponents of massive modularity is that human cognition is far more constrained, and far less flexible, than we have typically assumed. This too needs to be addressed, as it needs to be demonstrated that they are getting some significant set of the empirically available phenomena wrong. I think that they probably are—but in a book, this would have to be addressed in a great deal more detail. I think that it's probably right that they haven't got enough of a story about learning, and enough of an account of inter-modular control. That's the key take home of this chapter and I think that it's right. But the critiques could probably be softened, and the focus could be directed toward the construction of the positive account of control, as this is where the real action of the book will be, and focusing on the construction of a positive project will help to make this project useful to nativists and non-nativists alike—which I think that it should be)

- In chapter 4, the discussion of Dennett should be updated to include his more recent thoughts, which are anchored through the predictive processing framework. He is not explicit about seeing the problems that are raised here, but the fact that he has moved toward that framework might suggest that he too sees this kind of change as necessary. Additionally, there is a paper that just came out by Rob Goldstone and Georg Theiner that I think it would be super useful to read and respond to. This wasn't out at the time that the thesis was written, but there are multiple points of convergence between the ideas that they are trying to develop in discussions of group-level behaviour, and the kinds of claims about cognitive architecture that are being discussed in this thesis. Finally, it would be useful to discuss iterative reprocessing models at some point. They are of a piece with many of the kinds of arguments that are being developed here, though they are not basal ganglia driven—but they do share quite a bit in common with the cluster of positions that are discussed in Chapter 5.

- In the book, I would drop the discussion of dual process views. They aren't really focused on cognitive architecture, and they pull in a very different direction from the kinds of neurophilosophy that are being advanced in the remainder of the thesis. In chapter 5, the focus of the discussion of predictive processing should also focus more directly on the active inference model, as this is the point where the rubber really hits the road with respect to questions about control and flexibility.

- In the book version, I think that there needs to be a richer discussion of learning with regard to the basal ganglia, as an enormous amount of work has gone on in this area. It wasn't super important to do this on p.124ff of the thesis. But when this becomes a book this is something that should be spelled out, analysed, and expanded. My recommendation would be to draw more heavily on the kinds of learning and decision making models that have been proposed by Read Montague and by Fiery Cushman. For the book, the main focus will need to be on Part III, and these kinds of discussions will help to flesh out the positive picture, in ways that will increase the impact of this project on discussions of cognitive architecture.

- Footnote 121 should also be more fully expanded in the book. It should at least become a long section, as this is one of the really interesting and novel points that is being developed in Part III

- Since the discussion of behavioural flexibility played a critical role in spelling out the limitations of previous approaches to cognitive architecture, it needs to play a much more pronounced role in the articulation of the positive project. I would tend to think that it would be helpful to lead with this, or at least frontload it, as the linking of flexibility and control is what makes this position really interesting and plausible.

- In the discussion of ACT-R, see the recent work by Felipe de Brigard, as it goes quite a ways toward a more distributed architecture, which sustains the kinds of stability that Anderson posits.

THE UNIVERSITY OF SYDNEY

- Finally, the discussion of democratic decision making in the conclusion should be moved up quite a bit, and they should be discussed in more detail. I am inclined to think that they should probably become a chapter of their own in the book version of this project, as there is a lot of recent work within the cognitive sciences that really seems to support the kind of project that is being defended here. As I noted above, the recent paper by Goldstone and Theiner will be really helpful here. But far more importantly, there is a wealth of literature on collective decision making in insects and fish, which seems to suggest that they are using something like a Condorcet algorithm, which sets thresholds for decisions on the basis of recently experienced situations. You get a much more plausible model of this process when you look at the way that ants, for example, adapt their decisions in light of speed-accuracy considerations, using something like a diffusion model of decision making. And this looks like the kind of argument that's being advanced here; and it does so in ways that don't require thinking about the foibles of individuals, and the ways that groups of humans tend to pull away from democratic decision making. My guess is that talking about this stuff, and linking it to the model of the basal ganglia, would add a pretty serious punch to the project by showing that these kinds of democratic processes can generate biologically successful behavioural patterns.

I have included quite a few comments here, but I would like to close by noting that I really enjoyed reading this thesis, and I think that there are a lot of interesting and provocative ideas throughout. I really do hope that it is developed into a book project. I think it will have a big impact!

If your recommendation is for options 2 – 5, please provide detailed information relating to your recommendation in the appropriate box below. Only one section should be completed.

2) If the recommendation is to ***award with corrections***, please list below the required corrections.

3) If the recommendation is for ***revision and resubmission of the thesis,*** please provide a detailed list of your recommended errors and deficiencies that the student is required to address before the thesis can be re-examined.

4) If the recommendation is for ***award to another degree,*** please provide the reasons for this recommendation.

5) If the recommendation is for ***non-award,*** please provide reasons for this recommendation.

Please attach additional pages if required

Thank you for completing this report. You will be informed of the outcome of the examination once a decision has been made by the University.

THE UNIVERSITY OF SYDNEY

Higher Degrees by Research Examiner's report on thesis

Report due date: 16/05/2017

Name of examiner: Prof Robert Rupert

Examiner's institution: University of Colorado-Boulder

Name of Student: Ta Huang SID: 311057411

Faculty: Science Date of submission: 28/02/2017

Title of thesis: Neuro-Democracy: Self-Organization in the Embodied Mind.

Degree: PhD

A CONTENT AND PRESENTATION OF THESIS

As described in Clause 8 of Thesis and Examination of Higher Degrees by Research Policy 2015 , the thesis should:

a) be the student's own work, embodying the results of the work undertaken by the student during candidature;
b) form a substantially original contribution to the knowledge of the subject concerned;
c) afford evidence of originality by the discovery of new knowledge; and the exercise of independent critical ability;
d) form a cohesive and unified whole;
e) include a substantial amount of material that may be suitable for publication;
f) satisfactorily demonstrate that the student is able to identify, access, organise and communicate new and established knowledge;
g) be written in a standard generally acceptable to the discipline;
h) be written in English (except where permitted under the University of Sydney (Higher Degree by Research) Rule 2011)

B RECOMMENDATION

Please tick no more than one of the following recommendations:

After examination of the thesis and supporting material, I recommend that

(1) ***The student be awarded the degree without further conditions*** ..... ☐

or (2) ***The student be awarded the degree subject to corrections of the thesis to the satisfaction of the University***.. ☑
Corrections include errors or omissions in the thesis, such as incorrect citations, omissions, or typographical errors, which must be corrected but which do not alter the conclusions of the thesis, Corrections may also include additions or deletion of material in the text, tables, figures, or appendices. Changes should, in the opinion of the examiner, not require an additional period of research and should not result in the conclusions of the thesis being significantly altered. Corrections do not require the thesis to be returned to examiners, but can be adjudicated by the Chair of Examination.

or (3) ***The student not be awarded the degree, but be permitted to resubmit a revised thesis for examination following a further period of study***. ....☐
The thesis in its current form does not merit award. There are errors and/or deficiencies that, in the opinion of the examiner, substantially affect the argument or the conclusions of the thesis. Changes may include but are not limited to the provision of extra data or material. The student demonstrates sufficient ability that, after an additional period of study, a thesis of the required standard may be achieved. Following revision and resubmission, the thesis is re-examined.

I feel a further period of research extending over ... . months would be necessary.

Please indicate whether you would be willing to re-examine a revised thesis if so invited.... Yes ☐ No ☐

***Note**: This option is not available for a thesis that has already been revised and resubmitted for examination.*

or (4) ***The student not be awarded a doctoral degree but be awarded another degree for which they are eligible***...☐
The thesis is not considered satisfactory for the award of the degree for which it was submitted, but another degree for which the student is eligible may be awarded instead.

or (5) ***The student not be awarded the degree***..... ...☐
The thesis does not merit award of the degree and does not demonstrate sufficient ability by the student for a resubmitted thesis to achieve this merit. For example, the hypothesis and methods may be fatally flawed, therefore rendering the conclusions completely invalid and not capable of being rectified by an additional period of study.

C RELEASE OF EXAMINER'S NAME AND COMMENTS TO THE STUDENT

THE UNIVERSITY OF SYDNEY

I agree that the University may release my name and report to the student during or after the examination or as required by legislation. [In exceptional circumstances, the University may withhold information about the examiner]

Examiner name Robert D. Rupert 5/16/17 Date

Signature May 16, 2017

D GROUNDS FOR RECOMMENDATION

Please complete the following sections. In cases where examiners are not unanimous in their recommendation, this information will form the basis of the decision made by the University on the award or non-award of the degree.
Please note that the following boxes will expand according to the amount of text you type.

Is the ***content and presentation*** of the thesis consistent with the description in Section A of this form?

Yes ☑ No ☐

Please state below the grounds on which you base your recommendation. You may also wish to provide suggestions for the next steps in research or improvements for publication that are not required for the award of the degree.

The thesis clearly meets the standards (a)–(h). It is well written and organized; it represents an wealth of research on the part of the student; and, taken as a whole, it expresses a coherent, original, positive view of cognitive architecture and its component mechanisms responsible for flexible problem-solving by coherent agent. I recommend that the student vigorously pursue publication of this work, particularly the positive account developed in the later chapters (HECA with both distributed and centralized control and the mechanisms implementing that). A handful of first-rate papers could easily come out of these chapters (articulating and arguing for HECA, on the role of BG in cognitive control, on the relationship between BG and widely studied "executive control" processes in pre-frontal cortex, on the role of LCA's in a demon architecture, on the way in which the BG's control-structure – guided by reinforcement learning and the implementation of a restricted jury theorem – can support problem-solving in novel situations, and more). Although it can be seen as risky with regard to career strategy, I nevertheless encourage the student to turn quickly to the production a book manuscript that fleshes out and further his positive view. To be clear, the current document lacks some of the detail and polish that would be required for publication with a top press. But, the student has identified the basic components of a novel positive picture and, more importantly, how they interconnect. The student's resulting picture displays a deep unity, binding together a number of provocative smaller-scale points and proposals into a grander theoretical structure with significant promise as a solution to the deepest problems in philosophy of cognitive science (how to explain flexibility and coherence – hallmarks of intelligent systems). As such, I think this material will be most effectively presented in book-length form – perhaps preceded by a preliminary, though relatively comprehensive, theoretical statement in a venue such as Journal of Philosophy, Behavioral and Brain Sciences, or Synthese. In this way, its full scope and power can be appreciated. In pursuing this project I especially encourage the student to attend carefully to the relationships between the mechanistic and neural details (and models of them) and the philosophical claims; it can be difficult to convince readers of such connections, so, as much as possible, arguments for them must be slow, careful, and thorough.

It's less clear to me that the critical chapters will contribute substantively to the philosophical debate. That being said, I think there's potential for publication of the critical work on Massive Modularity Thesis (MM), in particular, the student's argument that MM's nativist commitment stands in the way of any MM-based account of the human ability to solve problems in novel contexts. The student will, however, have to grapple with the objection he sets aside in footnote 64 (and the generalization of the objection that I included in my comments inserted into the document itself).

On a related note, when the student engages critically with other views or contrasts his own view with competitors, I encourage the student to take more care to understand them charitably. To be more specific, the student should try to interpret the classical tradition in cognitive science (and not only Fodor's version of it) more accurately and charitably; rather than adopting a summary account formulated by critics (who frequently are setting up a straw view, just to knock it down – cf. Hurley on the sandwich model), it's better to go to the sources themselves; derive the characterization of the Classical Cognitivist Tradition directly from the work of Newell and Simon, Marr, Chomsky, Miller, Anderson, Pylyshyn, et al., else criticisms of it are likely to seem shallow or to appeal only to existing converts. Similar points: consider more fairly the empirical support for production systems (and more generally, Anderson's picture of the role of the BG), the supposed deep divide between embodied views and classical views (why can't a typical claim coming out of the embodied camp be Ramsified or modeled computationally?), and questions about implementation v. algorithmic level (won't someone who thinks like Fodor and Pylyshyn 1988 say that you've failed to capture the content of the messages passed during control processes involving BG, by failing to give a focused account of the algorithmic-level processes and of the kinds of representations that would participate at that level).

THE UNIVERSITY OF SYDNEY

If your recommendation is for options 2 – 5, please provide detailed information relating to your recommendation in the appropriate box below. Only one section should be completed.

2) If the recommendation is to ***award with corrections***, please list below the required corrections.

I've sent the .pdf with "sticky notes" inserted, marking typos and a missing reference (Clark 2001). Sticky notes include substantive comments as well (many of which make an appearance in the text that I entered into the preceding box), but responses to those comments are not required for the awarding of the degree.

3) If the recommendation is for ***revision and resubmission of the thesis,*** please provide a detailed list of your recommended errors and deficiencies that the student is required to address before the thesis can be re-examined.

4) If the recommendation is for ***award to another degree,*** please provide the reasons for this recommendation.

5) If the recommendation is for ***non-award,*** please provide reasons for this recommendation.

Please attach additional pages if required

Thank you for completing this report. You will be informed of the outcome of the examination once a decision has been made by the University.

THE UNIVERSITY OF SYDNEY

Higher Degrees by Research Examiner's report on thesis

Report due date: 16/05/2017

Name of examiner: Professor Timothy Schroeder

Examiner's institution: Rice University

Name of student: Ta Huang SID: 311057411

Faculty: Science Date of submission: 28/02/2017

Title of thesis: Neuro-Democracy: Self-Organization in the Embodied Mind.

Degree: PhD

A CONTENT AND PRESENTATION OF THESIS

As described in Clause 8 of Thesis and Examination of Higher Degrees by Research Policy 2015 , the thesis should:

a) be the student's own work, embodying the results of the work undertaken by the student during candidature;
b) form a substantially original contribution to the knowledge of the subject concerned;
c) afford evidence of originality by the discovery of new knowledge; and the exercise of independent critical ability;
d) form a cohesive and unified whole;
e) include a substantial amount of material that may be suitable for publication;
f) satisfactorily demonstrate that the student is able to identify, access, organise and communicate new and established knowledge;
g) be written in a standard generally acceptable to the discipline;
h) be written in English (except where permitted under the University of Sydney (Higher Degree by Research) Rule 2011)

B RECOMMENDATION

Please tick no more than one of the following recommendations:

After examination of the thesis and supporting material, I recommend that

(1) ***The student be awarded the degree without further conditions*** .... ☒

or (2) ***The student be awarded the degree subject to corrections of the thesis to the satisfaction of the University***.. ☐
Corrections include errors or omissions in the thesis, such as incorrect citations, omissions, or typographical errors, which must be corrected but which do not alter the conclusions of the thesis, Corrections may also include additions or deletion of material in the text, tables, figures, or appendices. Changes should, in the opinion of the examiner, not require an additional period of research and should not result in the conclusions of the thesis being significantly altered. Corrections do not require the thesis to be returned to examiners, but can be adjudicated by the Chair of Examination.

or (3) ***The student not be awarded the degree, but be permitted to resubmit a revised thesis for examination following a further period of study***. ..... …☐
The thesis in its current form does not merit award. There are errors and/or deficiencies that, in the opinion of the examiner, substantially affect the argument or the conclusions of the thesis. Changes may include but are not limited to the provision of extra data or material. The student demonstrates sufficient ability that, after an additional period of study, a thesis of the required standard may be achieved. Following revision and resubmission, the thesis is re-examined.

I feel a further period of research extending over ... . months would be necessary.

Please indicate whether you would be willing to re-examine a revised thesis if so invited.... Yes ☐ No ☐

Note*: This option is not available for a thesis that has already been revised and resubmitted for examination.*

or (4) ***The student not be awarded a doctoral degree but be awarded another degree for which they are eligible***…☐
The thesis is not considered satisfactory for the award of the degree for which it was submitted, but another degree for which the student is eligible may be awarded instead.

or (5) ***The student not be awarded the degree*** ..... …☐
The thesis does not merit award of the degree and does not demonstrate sufficient ability by the student for a resubmitted thesis to achieve this merit. For example, the hypothesis and methods may be fatally flawed, therefore rendering the conclusions completely invalid and not capable of being rectified by an additional period of study.

C RELEASE OF EXAMINER'S NAME AND COMMENTS TO THE STUDENT

THE UNIVERSITY OF SYDNEY

I agree that the University may release my name and report to the student during or after the examination or as required by legislation. [In exceptional circumstances, the University may withhold information about the examiner]

Examiner name Timothy Allan Schroeder ..... 21 May 2017 ..... Date

Signature……TS……..

D GROUNDS FOR RECOMMENDATION

Please complete the following sections. In cases where examiners are not unanimous in their recommendation, this information will form the basis of the decision made by the University on the award or non-award of the degree. *Please note that the following boxes will expand according to the amount of text you type.*

Is the ***content and presentation*** of the thesis consistent with the description in Section A of this form?

Yes……..☒ No……..☐

Please state below the grounds on which you base your recommendation. You may also wish to provide suggestions for the next steps in research or improvements for publication that are not required for the award of the degree.

The dissertation is a very nice and original piece of work on cognitive architecture. I admired both the conceptual clarity of seeing the possibility for a central but minimal action selection system and also the interpretive skill in taking the scientific literature on the basal ganglia and seeing its significance to more abstract debates. I'm not 100% sure I agree with all of the interpretations of the science, but by its nature cutting-edge scientific work is always open to such questions, and that is no serious fault of the dissertation.

If your recommendation is for options 2 – 5, please provide detailed information relating to your recommendation in the appropriate box below. Only one section should be completed.

2) If the recommendation is to ***award with corrections***, please list below the required corrections.

3) If the recommendation is for ***revision and resubmission of the thesis,*** please provide a detailed list of your recommended errors and deficiencies that the student is required to address before the thesis can be re-examined.

4) If the recommendation is for ***award to another degree,*** please provide the reasons for this recommendation.

5) If the recommendation is for ***non-award,*** please provide reasons for this recommendation.

Please attach additional pages if required

Thank you for completing this report. You will be informed of the outcome of the examination once a decision has been made by the University.

ABSTRACT OF THE THESIS

Neurodemocracy:
Self-Organization of the Embodied Mind

By

Linus Ta-Lun Huang

Thesis Supervisor:
Dominic Murphy

This thesis contributes to a better conceptual understanding of how self-organized control works. I begin by analyzing the control problem and its solution space. I argue that the two prominent solutions offered by classical cognitive science (centralized control with rich commands, e.g., the Fodorian central systems) and embodied cognitive science (distributed control with simple commands, such as the subsumption architecture by Rodney Brooks) are merely two positions in a two-dimensional solution space. I outline two alternative positions: one is distributed control with rich commands, defended by proponents of massive modularity hypothesis; the other is centralized control with simple commands. My goal is to develop a hybrid account that combines aspects of the second alternative position and that of the embodied cognitive science (i.e., centralized *and* distributed controls with simple commands). Before developing my account, I discuss the virtues and challenges of the first three. This discussion results in a set of criteria for successful neural control mechanisms. Then, I develop my account through analyzing neuroscientific models of decision-making and control with the theoretical lenses provided by formal decision and social choice theories. I contend that neural processes can be productively modeled as a collective of agents, and neural self-organization is analogous to democratic self-governance. In particular, I show that the basal ganglia, a set of subcortical structures, contribute to the production of coherent and intelligent behaviors through implementing "democratic" procedures. Unlike the Fodorian central system—which is a micro-managing "neural commander-in-chief"—the basal ganglia are a "central election commission." They delegate control of habitual behaviors to other distributed control mechanisms. Yet, when novel problems arise, they engage and determine the result on the basis of simple information (the votes) from across the system with the principles of neurodemocracy, and control with simple commands of inhibition and disinhibition. By actively managing and taking advantage of the wisdom-of-the-crowd effect, these democratic processes enhance the intelligence and coherence of the mind's final "collective" decisions. I end by defending this account from both philosophical and empirical criticisms and showing that it meets the criteria for successful solution.

Declaration

I certify that the intellectual content of this thesis is the product of my own work. Assistance received in preparing this thesis and sources have been acknowledged.

Linus Ta-Lun Huang

For P.B.

Acknowledgements

Working on this research project has been one of the most transformative experiences of my life. I would like to first express my deepest gratitude to my supervisor, Dominic Murphy and *de facto* associate supervisor, Eric Schwitzgebel, for giving me the freedom to explore my own voice. Their faith in me and this project, as well as their constant advice and encouragement, was a crucial motivating force in the completion of this thesis.

I also want to thank Kim Sterelny, Chris Eliasmith, Paul Thagard, and Mark Sprevak for their generous support during my several long visits to the Australian National University, University of Waterloo, and University of Edinburgh. They welcomed me into their research group and provided me with valuable resources. They have made this nomadic research life an extraordinary and enjoyable experience.

My appreciation also goes to Glenn Carruthers, Colin Klein, Bryce Huebner, Liz Schier, Bernard Balleine, Peter Godfrey-Smith, Sheldon Chow, Colin Allen, Adina Roskies, Kevin Gurney, Tom Stafford, Ahmed Moustafa, Michael Frank, as well as participants of the MacNap reading group, the SANU meeting (in particular, Ivan Gonzalez-Cabrera, Liz Irvine, and Pierrick Bourrat), and the Edinburgh Mind and Cognition Group for helpful feedback on my research. It would have been a lonely journey without them.

Special thanks go to Caitrin Donovan, a good friend and knowledgeable colleague, whose help has seen me through crises in life. She also provided invaluable assistance copy-editing and proofreading this thesis. To my friends at the Unit for History and Philosophy of Science (especially, Anson Fehross, Maria Kon, Claire Kennedy, Georg Repnikov, Gemma Smart, Kevin Keith, and Ian Lawson), as well as co-organizers of Minorities and Philosophy (Kari Greenswag, Yarran Hominh, Omid Tofighian, and Louise Richardson-Self), thank you for your warm company.

Finally, I thank my family for their love and patience, for being there with me as the stabilizing force through the ups and downs, and for continuing to make life worth living.

Table of Contents

1

Introduction

1. Introduction

How does the human mind emerge from brain processes? Recent research in philosophy and the cognitive neurosciences has made significant progress in understanding the neural mechanisms responsible for specific functions, such as those involved in reasoning and motor action. However, it remains unclear how these diverse neural mechanisms control and coordinate one another to generate coherent and intelligent behaviors (Eliasmith, 2013). To solve this "control problem," classical cognitive science posits central cognitive mechanisms to coordinate between perceptual and motor mechanisms (Fodor, 1983). Proponents of embodied cognitive science, skeptical of the existence of central mechanisms and the distinction between perceptual and motor mechanisms, often assume that integrated sensorimotor processes can self-organize in the production of intelligent behaviors (Clark, 1998). While embodied cognitive scientists reject central mechanisms on sound theoretical and empirical grounds, they have not adequately addressed how self-organized coordination works. Without an answer to the control problem, we cannot understand how intelligence emerges out of the interactions of less intelligent sensorimotor processes, nor can we comprehend how coherent behaviors arise from internally fragmented neural systems. In other words, there is currently a gap between our piecemeal understanding of cognitive mechanisms and a holistic understanding of the human mind. My thesis will provide a better conceptual understanding of how the mind self-organizes control.

In Section 2, I will begin by articulating the control problem in more detail and motivating its importance for theories of cognitive architecture. Then, I will distinguish three conceptually distinct sub-problems: the problem of architecture, the problem of coherence, and the problem of intelligence. In Section 3, I will briefly review two prominent solutions that have emerged from classical cognitive science and embodied cognitive science, respectively; neither, I shall argue, have

successfully solved the control problem. However, these two solutions are not exhaustive; rather, they are merely two positions in a two-dimensional solution space. Having illuminated this solution space, I will review two other positions: one is employed by the massively modular architecture; the other (which I will refer to, for the sake of convenience, as "Position X") has not been explored in philosophy, but it is supported by emerging empirical literature in cognitive neuroscience. In Section 4, I will introduce my positive account of neurodemocracy, which is a hybrid account that combines aspects of both Position X and the solution offered by embodied cognitive science. I end this chapter with an outline and plan for the rest of the thesis.

2. The Control Problem and Cognitive Architecture

Cognitive architecture, which is "a general proposal about the representations and processes that produce intelligent thought" (Thagard, 2012, p. 50), was a central topic in late twentieth-century philosophy of cognitive science. The centrality of cognitive architecture is reflected in Fodor's work on the classical symbolic approach to cognitive science and faculty psychology (Fodor, 1975, 1983), in debates between connectionists and classicists (Churchland, 1989; Fodor & Pylyshyn, 1988), in the emergence of dynamicism and embodied cognitive science (Beer, 2000; Clark, 1998; Port & Van Gelder, 1995), as well as in evolutionary psychology's proposal of massive modularity (Buss, 2005; Carruthers, 2006; Pinker, 2005). The cognitive revolution made common currency of cognitivism, the view that complex behavior is produced by inner representational states and information processing. It also provided a promising way of overcoming Descartes' mind-body problem: it was thought that we may finally explain how something as intelligent and flexible as human behavior emerges from the operation of physical mechanisms. Naturally, philosophers interested in how the mind works were drawn to this new and blossoming subfield of cognitive science, and were eager to spell out the conceptual contours and philosophical implications of contending theories. The results have been fruitful (Thagard, 2012).

In the last decade philosophers have turned their backs on the endeavor of cognitive architecture; instead, they have focused their attention on issues concerning specific functional domains of the cognitive system—for example, the mechanisms that underpin vision and social cognition (Eliasmith, 2013). However, I believe it is time to revisit debates on cognitive architecture and general intelligence in light of new advances in science.

Instead of focusing on the nature of representations and their transformation, which was the main focus in the last wave of cognitive architecture literature, this thesis will investigate the problem of control. I will start by motivating the control problem as an important challenge for theories of cognitive architecture and general intelligence. Then, I will suggest that this more general problem can be partly analyzed into three sub-problem: the problem of architecture, the problem of coherence, and the problem of intelligence. With the control problem and its three aspects clearly defined, we will be in a better position to assess the solutions that are currently on offer, as well as any future proposals.

2.1. The Control Problem

Solving the control problem, much like solving the problems of perception and motor planning, is essential for any complex biological cognitive system. Examples of control include decisions over which neural mechanisms are employed, the manner in which representations are manipulated in a given context, what actions are to be performed, as well as the flow of information between neural mechanisms, etc. The control problem can be concisely defined in the following manner:

> **Control Problem**: the problem of how a cognitive system controls its component mechanisms, and utilizes its information flexibly and selectively, to generate context-appropriate behaviors in a wide range of situations.

Despite significant differences between theories of cognitive architecture, the position that the human cognitive system is composed of neural mechanisms with diverse functional roles has attracted widespread consensus. One prominent example of this is the thesis that cognitive systems are made up of domain-specific cognitive mechanisms called *modules*. However, one need not commit to the existence of modules (nor to the related idea of functional specialization) to agree with this broader consensus. One alternative is Michael L. Anderson's theory of neural reuse (M. L. Anderson, 2014, 2015), according to which the human cognitive system is composed of neural circuits that are "functionally differentiated," but not "functionally specialized."[1] Insofar as human cognitive systems are made up of a massive number of distributed neural components, the control problem needs to be solved for intelligent behavior to emerge from their coordinated activities.

The control problem is by no means new. Descartes (in)famously posited an immaterial soul with the capacity to control, and coordinate between, perceptual and motor neural mechanisms. It was this posit that he used to solve the control problem and in doing so explain intelligent behaviors. It is clear, however, that we can no longer accept a non-physical entity, mediating between neural mechanisms and conferring intelligence to behaviors, as a viable solution.

More "naturalistic" constructs, such as *control processes*, the *central executive*, and *supervisory attention system*, have been posited to solve the control problem (Monsell, 1996). However, many of these scientific constructs continue to function like a homunculus fiddling mysteriously behind the scenes. As Newell explains:

> A major item on the agenda of cognitive psychology is to banish the homunculus (i.e. the assumption of an intelligent agent (little man) residing elsewhere in the system, usually off stage, who does all the marvellous things that need to be done actually to generate the total behaviour of the subject). It is the homunculus that actually performs the control processes in Atkinson & Shiffrin's (1968) famous memory model, who still does all the controlled processing (including determining the strategies) in the more recent proposal of Shiffrin & Schneider (1977), who makes all the confidence judgements, who analyses all the payoff matrices and adjusts the behaviour appropriately, who is renamed the "executive" in many models (clearly a promotion); … (Newell, 1980, p. 715)

The control problem remains an active research topic in contemporary cognitive science, even if it is often referred to in different terms. For example, the input problem at the core of debates between Fodor and massive modularists (Fodor, 2000; Pinker, 1999), which concerns how information is routed to the right place at the right time, is one version of the control problem. Similarly, Murray Shanahan uses the term “coalition formation problem” to refer, essentially, to the same problem, even if he focuses more on the information, rather than the mechanisms, that

---

[1] The key difference between them is that, for functional differentiation, the same neural circuits “are used and reused in multiple cognitive/behavioral circumstances, across traditional task categories" (M. L. Anderson, 2014, p. 103), such as face perception or mind-reading. In fact, they serve across different tasks by participating in the different neural coalitions responsible for different cognitive functions. As a result, unlike theories of modularity in which a one-to-one mapping between modules and cognitive functions is posited, the neural reuse theory assumes a many-to-many mapping between neural circuits and cognitive functions. That is: “Neural reuse departs from interactive specialization by emphasizing the participation of neural elements in *multiple* coalitions” (M. L. Anderson, 2015, p. 7).

process it (Shanahan, 2012).[2] Finally, in M. L. Anderson's neural reuse theory, the control problem is addressed by positing active "neural search" processes, which are responsible for "the rapid testing of multiple neural partnerships to identify functionally adequate options" (M. L. Anderson, 2014, p. 58) so that "a diverse behavioral repertoire is achieved via the search for and consolidation of multiple, nested, and overlapping neural coalitions" (M. L. Anderson, 2015, pp. 1–3).

In short, the control problem remains an important problem for any large-scale cognitive system. As a result, understanding how human cognitive systems solve this problem is vital for understanding human intelligence.

2.2. Three Sub-Problems of Control

The control problem needs to be approached on two fronts: architecture and decision-making. The problem of architecture concerns the "infrastructure" for coalition formation. Yet, given an adequate infrastructure for flexible coalition formation, control decisions still need to be made as to which neural components to incorporate into a coalition, when they should be incorporated, and how they should interact. In particular, there are two important criteria for a good control decision: first, it should be an intelligent decision that helps form neural coalitions that enable the agent to overcome environmental challenges, and second, it should be coherent with control decisions made at other loci in the cognitive system. The decision-making side of the control problem can thus be further analyzed into at least two sub-problems: the problem of coherence and the problem of intelligence.

Note that this is not an exhaustive list of all sub-problems on the decision-making side of the control problem.[3] I focus on these problems because, insofar as they pertain to two defining features of human behavior (i.e. intelligence and coherence), solving them is of the utmost importance to the endeavor of cognitive architecture. With this qualification in mind, I will now clarify the problem of architecture, the problem of coherence, and the problem of intelligence respectively.

The problem of architecture concerns how the connections between neural components of a cognitive architecture should be set up in order to support their flexible recruitment into "neural coalitions" that solve a wide range of environmental problems. This problem is challenging because to solve it optimally, the organization of neural connections needs to maximize flexibility while minimizing the costs that flexibility entails. On the one hand, connecting every neural component to each other would create a very flexible organization, but the total number of connections necessary to do so are too high to be biologically plausible for any large-scale cognitive system. On the other hand, if only a limited number of connections are built between neural components, the cognitive system may not be able to "draw upon the full battery of its neuronal resources, mixing and matching them as required, to find an effective, and sometimes innovative, response to unfamiliar situations" (Shanahan, 2012, p. 2707).

---

[2] According to Shanahan, when a cognitive system overcomes the coalition formation problem, the system achieves "cognitive integration," which is a state "when the animal brings the totality of what it knows to bear on the ongoing situation—its grasp of the sensorimotor contingencies of multiple domains and its understanding of the associated affordances, plus the full contents of both its long-term (episodic-like) memory and its short-term (working) memory" (Shanahan, 2012, p. 2704).

[3] That is, I do not assume that the decision-making side of the control problem will be completely solved when the two sub-problems discussed here are solved.

The problem of architecture is not the main concern of this thesis, though I will touch on it briefly as we discuss the various solutions to the control problem—as we will see, the current solutions offered by classical and embodied cognitive sciences have largely solved it. Instead, our primary focus will be the two sub-problems that relate to the decision-making side of the control problem: the problem of coherence and the problem of intelligence. I will now discuss each in turn.

The problem of coherence concerns the challenge of making coherent control decisions in order to generate large-scale, goal-directed behaviors. This problem is challenging because numerous control decisions need to be made correctly at various loci of a cognitive system at any given moment so that (1) adequate neural coalitions can be formed to solve the problems the cognitive system is facing, and (2) neural coalitions, that have been established for different purposes, do not conflict with one another to the degree that completely undermine one another's goals. Due to the limited scope of this thesis, I will only be concerned with synchronous coherence and will bracket the issue of asynchronous coherence, which pertains to control decisions made at different times. My reason for doing so is that asynchronous control is less fundamental than, and presupposes, synchronous coherence.

The problem of intelligence concerns the means by which intelligent control decisions emerge from the interaction of less intelligent neural components. To appreciate the difficulty of making intelligent control decisions, we need to recognize the flexibility of human cognitive systems; human beings are capable of solving a wide range of novel problems in open-ended environments. As a result, neural components that are functionally specialized or differentiated need to be assembled into neural coalitions that have the capability to "move beyond the tried-and-tested, to transcend domain-specific expertise" (Shanahan, 2012, p. 2707).[4] That is, intelligent control decisions need to be made in various loci of the cognitive system to produce neural coalitions that meet novel environmental challenges. However, it is not clear how the competence necessary for making intelligent control decisions can emerge from the interaction of neural components which themselves do not already possess such competence.

In conclusion, the problems of architecture, coherence, and intelligence are three important sub-problems of control that need be addressed by any large-scale cognitive system that is composed of numerous distributed components. As a result, understanding how human cognitive systems solve these problems is paramount to the research program of cognitive architecture. In the next section, I will turn to the conceptual space for the control problem's solution.

3. Solution Space for the Control Problem

In this section, I will begin by introducing two prominent paradigms in cognitive science—classical and embodied cognitive science—as well as the solution they each offer to the control problem. I will then analyze the core features of these solutions: briefly put, classical cognitive science utilizes centralized control and a rich message-passing strategy, while embodied cognitive science relies on distributed control with a simple message-passing strategy. Articulating these core features discloses the solution space as two-dimensional: the first dimension represents control strategies ranging from centralized to distributed, while the second dimension represents message-passing strategies ranging from rich to simple. Plotting the two prominent positions in this two-dimensional space

[4] Neural components become functionally specialized or differentiated through a combination of innate or developmental factors.

also reveals that there are at least two other positions available. I end section 3 by briefly introducing them.

3.1. Classical Cognitive Science and Its Solution to the Control Problem

The classical paradigm has not only enjoyed historical predominance in cognitive science, but it also continues to remain a mainstream view of the mind. It is associated with several models of cognition and cognitive architecture, including the model Fodor advocates in *The Modularity of Mind* (Fodor, 1983), Anderson's "Adaptive Control of Thought—Rational" architecture (J. R. Anderson, 2007), and the library model of cognition (Samuels, 1998). These associated cognitive architectures and models are often referred to as "classical sandwich" models (Hurley, 2001). This name aptly captures one of their key features: central cognitive mechanisms (in the narrow sense of mental mechanisms responsible for modeling the world and planning) are segregated from, and 'sandwiched' between, perceptual and motor mechanisms. In this section, I will introduce three main features of the classical sandwich models. Having done this, I will discuss the solution these models offer to the control problem, including two core characteristics that this solution exhibits. I will, however, postpone my evaluation of this solution until Chapter 2, after I discuss in detail the Standard Account, a prototypical model of classical cognitive science.

According to Hurley, there are three characterizing features of classical sandwich models. First, perceptual and motor mechanisms are separate and peripheral. Second, central cognitive mechanisms, which interface with perceptual and motor mechanisms respectively, are the source of human intelligence. Finally, the central cognitive mechanisms have the classical computational architecture "at the right level of description" (Hurley, 2001, p. 3).[5]

The first two features jointly constitute what Hurley calls "vertical modularity" (Figure 1.1). In the classical architecture, the mind decomposes into perceptual, central, or motor modules.[6] Information processing flows from perceptual to central modules, then to motor modules in a primarily linear or one-way manner. Perceptual modules function to extract information from inputs of different sensory domains to produce perceptual representations. Central modules are where belief-updating and deliberation happens, and both depend on the construction and manipulation of detailed, internal representations of the world. Once a decision is made in the central modules, a command is sent to motor modules to program the movements to be executed.

The third feature of classical computational architecture concerns central modules, and is motivated by several assumptions about the capacity of human thought such as systematicity and productivity, as well as presumed isomorphism between the subpersonal processes underlying rational thoughts

---

[5] A classical computational process is a computational process with some distinctive characteristics. First, the states being manipulated are symbolic representations that have a combinatorial syntax and semantics. They belong to a representational system where (a) molecular (structurally complex) representations are built up systematically (syntactically-structured) out of atomic representations, and (b) the content of a molecular representation is a function of its constitutive atomic representations and its syntactic structure. Second, the computational process is defined over, and causally sensitive to, the syntactic properties of the representations. Although the symbolic representations have both semantic and syntactic properties, the computational system is only responsive to, and manipulates them according to, their syntactic properties. For more details, see (Haugeland, 1985).

[6] The concept of modularity here is a minimal one, which refers to functionally characterized computational mechanisms. As a result, central systems in Fodor's model of the mind are also modules under this definition.

and their conceptual structures at the personal level (Hurley, 2001).[7] The personal-level patterns that exist in human thoughts and behaviors, such as the structures of rational thoughts, are assumed to be mirrored directly in the sub-personal processes in the central modules. In other words, rational thoughts and the relevant competence for rational thought are *not* considered to be emergent phenomena, even if (theoretically) emergent patterns, effects, and capacities can arise from the complex interactions of classical computational processes.[8]

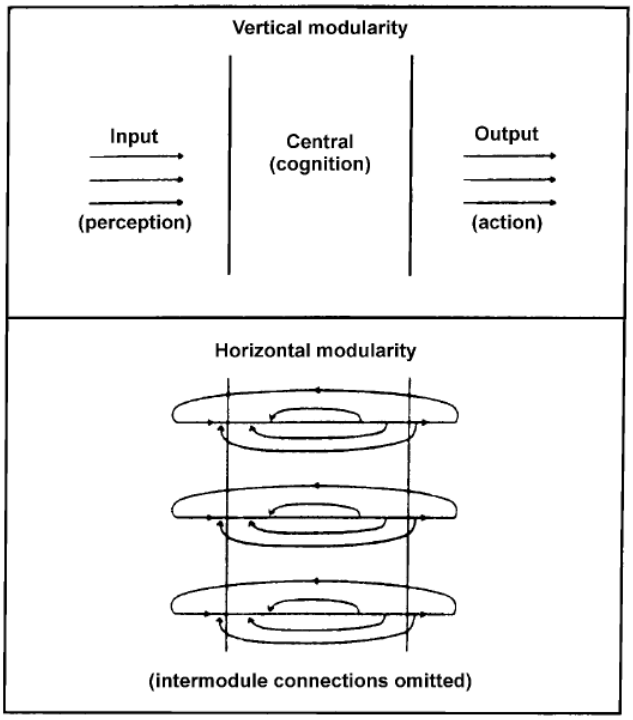

Figure 1.1 Vertical modularity and horizontal modularity. Reprinted from (Hurley, 2001).

Central modules are the classical sandwich models' solution to the control problem (Figure 1.2). According to Andy Clark (1998), this solution has two characteristics: First, the classical sandwich models depend on centralized control strategy. On the one hand, central modules (e.g., the Fodorian central systems) are considered to be the locus of human intelligence and possess the competence to coordinate other modules. They receive information from everywhere in the cognitive system and use this information to build and update detailed models of the world. Based on their knowledge of the world, the central modules can make flexible decisions as to what is best to do in order to meet their goals in different environments. On the other hand, perceptual and motor modules are relatively dumb, inflexible mechanisms that cannot learn to perform novel functions and have "their job descriptions carved in stone" (Dennett, 1991, p. 260). That is, perceptual and motor modules are what Dennett calls the "bureaucratic" agents in a "bureaucratic hierarchy" because they are under the strict control of the intelligent, higher-level central modules, which "dictate which bit of means/ends analysis they are authorized to perform" (Dennett, 1991, p. 236). For example, the motor modules are only authorized to work out the best motor programs to carry out the more abstract plans determined by the central modules, but have no authority to determine the plans themselves. In short, we can think of the centralized control strategy as casting

[7] For a detailed discussion of the isomorphism assumption and its problems, see (Huebner, 2014, Chapter 3).

[8] We will discuss the concept of emergence in more depth in the next section.

the central module in the role of neural commander-in-chief: the commander-in-chief receives information about various situations from his subordinate bureaucrats, then uses it to deliberate and produces commands for subordinate bureaucrats to execute.

Second, the classical sandwich models depend on a rich message-passing strategy. A rich message-passing strategy, according to Clark (1998), has two features: it uses representations of rich contents in general-purpose/amodal format to control other cognitive mechanisms,[9] and depends both on rich internal models, and the complex construction and transformation of representations for their control function. Central modules (e.g., the central systems) utilize the rich message-passing strategy for their control function, as they depend on the complex construction and transformation of rich internal models for belief-updating/planning; they also communicate with perceptual and motor modules with the detailed, amodal representations required by classical computationalism.

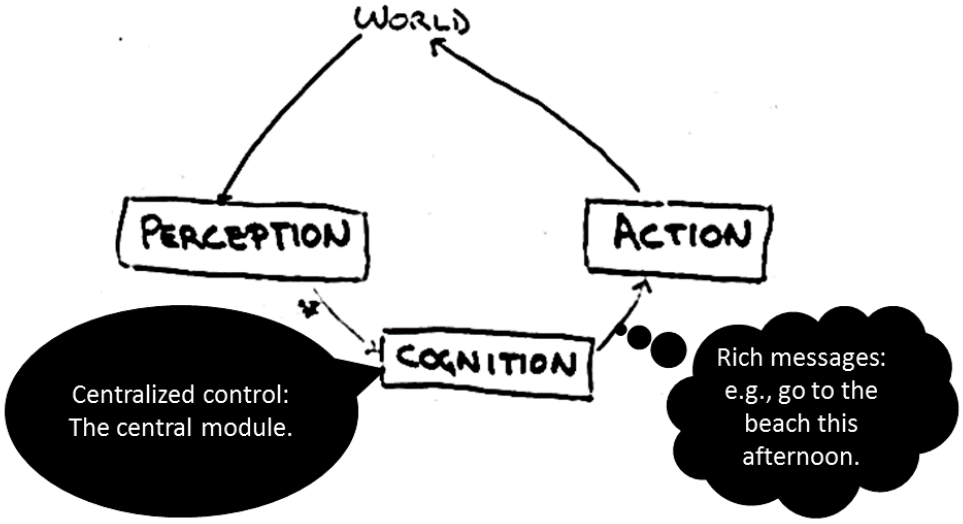

Figure 1.2 Classical cognitive science's solution to the control problem.

In short, classical cognitive science's solution to the control problem is central modules, which utilize centralized control with a rich message-passing strategy. I will discuss the paradigmatic model of classical cognitive science, the Standard Account, in more detail in Chapter 2. As we will see, the Standard Account, were it true of human cognitive systems, would provide an excellent solution to the control problem. However, fatal attacks on both empirical and theoretical fronts have rendered the Standard Account an implausible model of the human mind.

3.2. Embodied Cognitive Science and Its Solution to the Control Problem

What I refer to here as "embodied cognitive science"' is a broad theoretical framework that not only includes 4E cognition—that is, embodied, embedded, extended, and enactive cognition[10]—but also

[9] Because of the need to coordinate mechanisms of different domains using rich messages, there is pressure to depend on representations in general-purpose formats (Clark, 1998).

[10] Very briefly, according to the embodiment thesis, cognition depends on not just the brain, but also the body. According to the embedding thesis, cognition routinely depends on structures in the natural and social environment. According to the extension thesis, the vehicle of cognition extends beyond the boundaries of individual organisms.

the earlier society of mind framework (Dennett, 1991; Minsky, 1986). As such, embodied cognitive science is a diverse set of theories connected through some common assumptions, including a rejection of the individualistic Cartesian image of mind that is characteristic of classical cognitive science. In the following, I will first introduce embodied cognitive science by focusing on two features that are relevant to this thesis: first, the embodiment thesis, and second, the abandonment of three key features of classical cognitive science. Having done this, I will discuss the solution to the control problem provided by embodied cognitive science and its two characteristics. Finally, I will provide a brief evaluation of this solution in terms of its capacity to deal with the problems of coherence and intelligence.

Embodied cognitive science conceives of cognition as "representationally distributed, computationally dynamic, and as properly characterized only by reference to details of bodily realization" (Wilson & Foglia, 2011, p. 11). Specifically, the embodiment thesis plays a central role in embodied cognitive science.[11] A general formulation of this thesis can be stated concisely as follows:

> **The Embodiment Thesis**: Many features of cognition (meant here in the broad sense to include both central cognition and sensorimotor information processing) are dependent, causally or constitutively, on the sensorimotor parts of the brain and the (non-neural) body of an agent (Wilson & Foglia, 2011).[12]

Importantly, for the purposes of this thesis, embodied cognitive science rejects the three main features of classical cognitive science discussed earlier in Section 3.1. First, *contra* classical cognitive science, cognitive mechanisms in an embodied architecture do not separate into peripheral perceptual and motor "vertical modules," with central modules mediating between them. Instead, unified sensorimotor "horizontal modules" function without the mediation of central cognitive modules (Hurley, 2001). As Clark puts it, "real-time, real-world success is no respecter of this neat tripartite division of labor. Instead, perception is itself tangled up with specific possibilities of action—so tangled up, in fact, that the job of central cognition often ceases to exist" (Clark, 1998, p. 51). Moreover, instead of seeing the casual flow between perception and action as primarily linear and one-way and their relation as merely instrumental (i.e., one is merely a means to the other, and vice versa), we should see the functions of perception and action as realized by dynamical, sensorimotor mechanisms with circular causal flows from perception to action and vice versa, often with the environment playing an important mediating role. In other words, embodied cognitive science adopts an account of the mind "in which perception and action co-depend on dynamically circular subpersonal relations and as a result may be more than merely instrumentally interdependent" (Hurley, 2001, p. 3).

Embodied cognitive science also rejects the idea that central modules are the source of intelligence. In an embodied cognitive architecture, there are no intelligent central modules mediating between,

---

According to the enactment thesis, "the experienced world is portrayed and determined by mutual interactions between the physiology of the organism, its sensorimotor circuit and the environment" (Wilson & Foglia, 2011).

[11] Some versions of the embedding, extension, and enactment theses require the embodiment thesis because exploiting the environmental structure (the embedding thesis), integrating external entities into the functioning of cognitive systems (the extension thesis), and enacting one's experience (the enactment thesis) rely heavily on the sensorimotor capacities of organisms (Menary, 2010; Robbins & Aydede, 2008).

[12] I will spell out this more general embodiment thesis into several specific claims in Section 3.1 in Chapter 4.

or dictating to, the relatively dumb perceptual and motor modules. There are only sensorimotor mechanisms, which produce intelligent behavior by coordinating with one another.

Finally, embodied cognitive science rejects the idea that central modules are implemented, isomorphically, in a classical computational architecture. Instead, the mechanisms and competences underlying central cognition (and cognition in the broader sense) are implemented by action-centric, modality-specific representations and transformations, instead of the amodal and action-neutral representations and transformations required by classical computational architecture. Importantly, because they abandon the idea that a central module serves as the source of human intelligence, human intelligence must instead emerge out of the interactions of sensorimotor mechanisms that are semi-intelligent and semi-learning-capable to varying degrees. That is, both intelligent behaviors and the relevant competences are emergent phenomena.[13]

Embodied cognitive science's solution to the control problem is, unsurprisingly, the inverse of that proposed by classical cognitive science. The solution replaces the classical metaphor of neural commander-in-chief with that of "pandemonium competition": intelligent and coherent behavioral patterns emerge from the interactions of neural "agents" (as well as bodily and environmental mechanisms).[14]

The embodied solution has two key characteristics (Figure 1.3). The first is distributed control: instead of assigning control to a single intelligent neural mechanism which is privy to all information and responsible for coordinating other neural mechanisms, control is distributed across neural mechanisms. The second is a simple message-passing strategy: neural mechanisms use messages "whose content rarely exceeds signals for activation, suppression, or inhibition" for coordination (Clark, 2001, p. 140). Neural mechanisms also tend to rely on specialized internal models, which operate on highly proprietary or special-purpose formats for their control function; when neural mechanisms coordinate using simple messages alone, there is less pressure for them to share the representational format of their internal models (Clark, 1998).

Rodney Brooks' subsumption architecture is a simple illustration of how an embodied cognitive architecture solves the control problem to generate coherent and intelligent behavior. The subsumption architecture is composed of "horizontal modules" or "layers" (Figure 1.3), each of which constitutes a self-contained pathway that dynamically loops through internal sensorimotor mechanisms as well as the environment, and is capable of performing specific tasks by itself (Hurley, 2001, p. 7).

Take, for example, a simple robot's cognitive system composed of three layers (Figure 1.3) (Clark, 1998): Layer 1 produces exploratory behavior by generating random movement so that the robot can wander around the environment. Layer 2 is responsible for more goal-directed movement,

---

[13] Emergent phenomena, according to Clark's characterization (Clark, 1998, 2001), are higher-level effects, patterns, or capacities that are the results of a certain class of complex (e.g., non-linear, circular, temporally asynchronous, etc.) interactions among lower-level components of the system. As a result, emergent phenomena are phenomena that can be best understood in terms of the changing values of collective variables— i.e., variables that track higher-level patterns that result from complex interactions among multiple lower-level elements. Note that the higher-level vs. lower-level relations here refer to the part-whole relations between compositional levels (Wimsatt, 1994, p. 222).

[14] Clark thinks there are (at least) three categories of control solutions: the global dissipative effects of neuromodulators, the external scaffolding of natural, social, and technological environments, and the internal signaling of neural control structures (Clark, 2001, p. 100). I only focus on the internal signaling strategy in my thesis, but I believe the former two play important roles in creating intelligent and coherent behaviors as well.

which it initiates by determining a location and driving the robot towards it. Layer 3 manages obstacle avoidance by stopping the robot if an obstacle is ahead and reorienting it towards an unblocked direction.

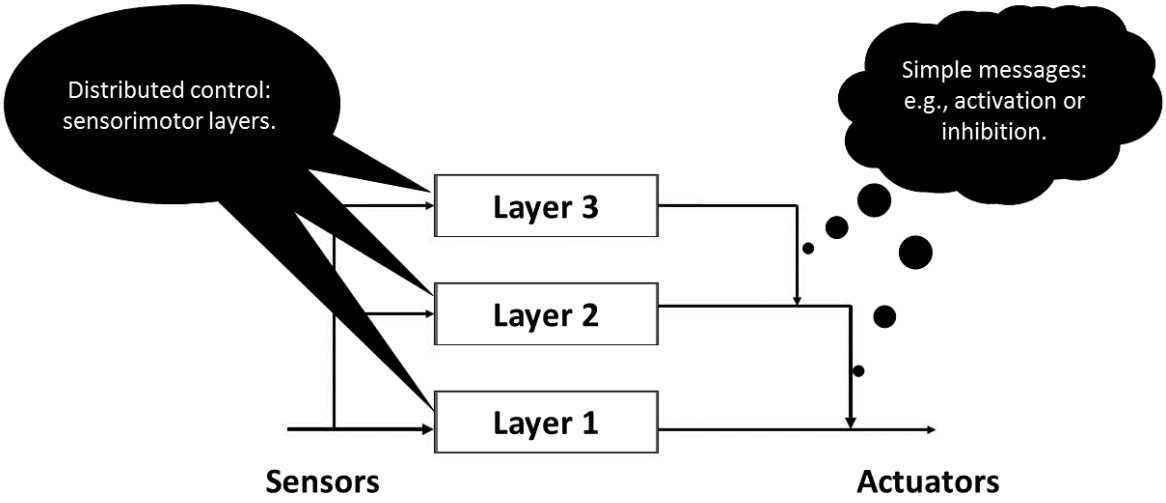

Figure 1.3 Embodied cognitive science's solution to the control problem as illustrated in a subsumption architecture.

The three layers need to be coordinated in order for the robot to exhibit intelligent and coherent behavior. Rather than relying on the intervention of an intelligent central module, their coordination depends on two design features. The first is "carefully orchestrated couplings" between each layer and selected features of the environments (Clark, 1998, p. 31): Layer 1 is activated automatically in cases where there are no interesting objects present in the environment; Layer 2 is activated when an interesting object (say, food) is detected; and Layer 3 is activated by environmental obstacles. The second feature is carefully set-up inter-layer inhibition and facilitation: in this simple cognitive system, Layer 3—when activated—will inhibit Layer 2, while Layer 2—when activated—will in turn inhibit Layer 1. These two features jointly produce a simple robot that exhibits the somewhat intelligent and coherent behaviors of foraging and obstacle-avoidance.

In this short passage, Hurley nicely summarizes how she thinks human intelligence and rationality can emerge from the complex interaction of layers:

> Very crudely, some layers get turned on and others turned off, in a totality of ways that count as rational overall in the circumstances. On this view, rationality is a higher-order property of complex patterns of response, which emerges from the layers of direct dynamic couplings between organisms and their structured environments. (Hurley, 2001, p. 10)

The design features employed in the previous example, despite their appeal, are too simple to solve the control problem for a more complex cognitive system. Clark has identified an additional design feature in more complex cognitive systems, such as the human brain: the deployment of "neural control structures." Neural control structures are "neural circuits, structures, or processes whose primary role is to modulate the activity of other neural circuits, structures, or processes" (Clark, 1998, p. 136). That is, instead of tracking and controlling external states of affairs, these structures track and control a cognitive system's inner economy by activating and inhibiting the layers. They do so, specifically, by integrating information from different layers, and modulating their interaction in order to generate more flexible and coherent behaviors. Note that these neural

control structures are nothing more than additional layers: they are self-contained pathways capable of performing a specific function by themselves, and they coordinate with other layers through simple messages of inhibition and activation. As Clark stresses:

> … there is no sense in which the [neural control structure] has access to all the information flowing through the system. [Neural control structures] are not executive controllers privy to all the information in the system, so much as simple switching agents, opening and closing channels of influence between a wide variety of inner processors and components. (Clark, 1998, p. 101)

To sum up, in the embodied cognitive architecture, control is not concentrated in an intelligent central module that has access to a large amount of intrasystem information, or depends on detailed internal models and the transfer of rich information for their control function. Rather, neural mechanisms form a distributed set of neural control structures, each of which has access to task-specific information, depends on highly proprietary internal models for their control function, and coordinates different task-specific mechanisms using simple messages that either activate or inhibit them.

Embodied cognitive science's solution to the control problem is theoretically plausible and well supported by recent empirical literature in decision-making and control (Botvinick & Cohen, 2014; Cisek & Kalaska, 2010; Glimcher & Fehr, 2014; Seth, Prescott, & Bryson, 2012). However, it remains unclear how it can address the problems of coherence and intelligence.[15] It remains to be seen, that is, how a massive set of fragmented layers and control structures—each of which has only partial information about the world—can give rise to coherent and intelligent behavior. As Clark puts it:

> There is something deeply fragmentary about the vision … of the natural roots of intelligent behavior in which efficient response depends on the presence of what the cognitive neuroscientist V.S. Ramachandran calls a "bag of tricks."… a mixed bag of relatively special-purpose encodings and stratagems whose overall effect is to support the particular needs of a certain kind of creature occupying a specific environmental niche. … How might large-scale coherent behavior arise from the operation of such an internally fragmented system? (Clark, 2001, p. 100)

A neural control structure, by integrating and extracting information from multiple sources, can modulate some of the layers in ways that make them behave more coherently. However, it is unclear how a massive set of distributed neural control structures can make coherent control decisions. That is, questions remain concerning what would, in turn, coordinate these neural

---

[15] The problem of architecture can be addressed, for example, by the connective core hypothesis proposed by Shanahan (2012). To introduce the hypothesis, I need to first clarify the concept of a hub node, which comes from the study of the connectome of the brain. Connectome is the large-scale organization revealed by analyzing anatomical pathways between brain regions using the mathematical theory of complex networks (graph theory). Roughly, a hub node is a brain region that has (relatively) a very large number of connections to other regions. As a result, a hub node plays a very important role for communication between different brain regions. A connective core is a small set of hub nodes that are (1) topologically central in the network of a cognitive system and (2) densely connected to each other. Specifically, Shanahan claims that three important functions are served by the connective core: broadcast of information, medium of coupling of different neural regions, locus of competition (such as process of control decision-making). Connective cores help solve the problem of architecture because " information and influence can funnel into and fan out from the connective core" in a way that supports the flexible coalition formation, and do so without the expensive "combinatory wiring" that connects every neural component to each other (Shanahan, 2012, p. 2709).

control structures and resolve their conflicts for the better rather than for the worse. Embodied cognitive science has not yet shown us, in other words, how to solve the problem of coherence. The problem of intelligence remains similarly unresolved: we still lack a plausible account of how an intelligent control decision can be generated by "pandemonium competition" among layers and neural control structures that possess only partial information about the world and depends on highly proprietary internal models.

Embodied cognitive science is an important alternative approach to cognition. It deals with the problem of control by positing distributed control structures utilizing a simple message-passing strategy. While this solution is theoretically plausible and empirically supported, it cannot—at least as it stands now—adequately address the problems of coherence and intelligence. In Chapter 4 and Chapter 5, I will examine new empirical developments in embodied cognitive science in order to evaluate the extent to which it has made progress on the control problem.

3.3. Two Dimensions of the Solution Space and Alternative Positions

If neither classical nor embodied cognitive science are able to provide us adequate solutions to the control problem, what options are we left with? Fortunately, these two dominant solutions do not exhaust all conceptual possibilities. Their defining characteristics—centralized control and a rich message-passing strategy on the one hand, and distributed control and a simple message-passing strategy on the other—are in fact conceptually independent from one another, respectively. As a result, the dominant solutions are merely two of at least four available positions in a two-dimensional solution space. In the following, I will describe this solution space and, in doing so, introduce the two alternative positions. The first position is adopted by the massively modular architecture, while the other, Position X, will form an important part of my positive account of neurodemocracy.

Even though centralized control is often paired with a rich message-passing strategy, these control features are not mutually entailing. That is to say, a centralized control strategy does not, as a matter of definition, require controllers to use rich messages to communicate with other mechanisms, nor does a rich message-passing strategy necessitate the involvement of central controllers for its execution. Likewise, the relationship between distributed control and a simple message-passing strategy is also a contingent one. As a result, we can conceptualize these characteristics along two dimensions of a solution space (Figure 1.4).

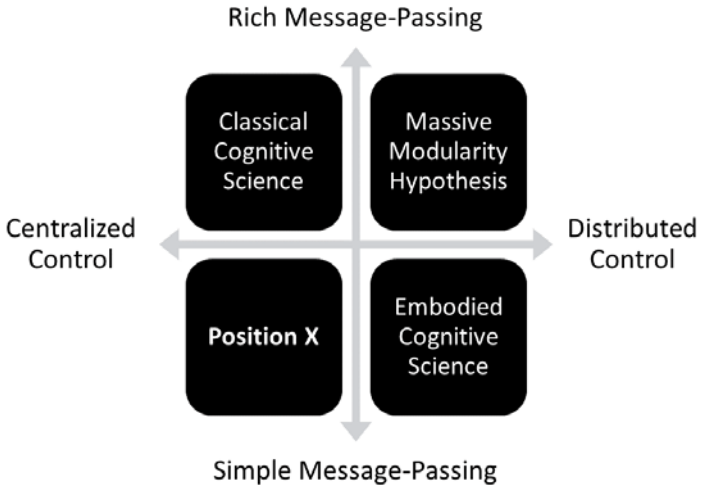

Figure 1.4 The solution space for the control problem.

The first dimension of the solution space, as depicted by the horizontal axis, concerns the centralization of the control mechanisms. The left-side of the horizontal axis represents the centralized control strategy, which posits central control mechanisms for controlling perceptual and motor mechanisms. On the other side is the distributed control strategy, which assumes that control emerges entirely from interactions among distributed neural components and control structures.

The second dimension of the solution space, as depicted by the vertical axis, concerns the "message-passing" strategy for control. On the rich message-passing strategy end of this dimension, we have control mechanisms that coordinate with representations with detailed content in a general-purpose code. These mechanisms rely heavily on complex internal models, as well as the construction and transformation of representations, for their control function. On the simple message-passing strategy end of this dimension, control mechanisms utilize control signals, whose representational contents consist merely of activation or inhibition, to encourage or discourage the activities of other neural components. These control mechanisms also tend to rely on simple internal models that operate on special-purpose formats.

Representing the classical and embodied solutions in this manner makes it apparent that there are at least two alternative positions available. The first of these uses a rich message-passing strategy without centralized control. This type of solution is adopted by some massive modularity models (Barrett, 2005; Carruthers, 2006; Sperber, 1994), which I will discuss in Chapter 2 and Chapter 3. According to these models, modules coordinate with one another using rich messages.[16] However, massive modularists posit not one, but multiple central modules that mediate between perceptual and motor modules. As a result, the precise means by which these (central and peripheral) modules coordinate with each other becomes a substantial issue. A satisfactory account will need to be provided regarding how these modules, through interacting with each other, can resolve the control problem so that the right module will handle information processing in the right context and the right time. As I shall argue in Chapter 3, these massive modularity models are not capable of addressing this issue and so fail to solve the control problem.

Finally, the fourth type of solution is Position X, which pairs centralized control with a simple message-passing strategy. As I shall argue in Chapter 6, Position X is not only conceptually possible, but is also empirically well-supported by literature on the basal ganglia—a group of subcortical neural structures—which has emerged in the last twenty years (Gurney, Prescott, & Redgrave, 2001a, 2001b; O'Reilly & Frank, 2006; Wiecki & Frank, 2013). However, an adequately developed philosophical account based on Position X still needs to be advanced. Demonstrating the plausibility of this position, which is a key component of my neurodemocracy account of control, will be the task of Chapter 6 and Chapter 7.

In this section, I have briefly discussed the dominant solutions offered by classical and embodied approaches to cognitive science. Analyzing their core features discloses the broader solution space they occupy, which includes two additional alternatives. One of these alternatives, Position X, will be an important component of my own positive account of neurodemocracy.

---

[16] The enzymatic routing architecture seems to allow modules to interact through inhibiting and activating each other (Barrett, 2005). However, signals exert inhibitory or excitatory effects on other modules based on their relatively rich representational content.

4. Neurodemocracy: A Hybrid Account

In this section, I will introduce the positive account that I will develop and defend in this thesis. Neurodemocracy is a hybrid account of control that incorporates aspects of both Position X and the solution favored by embodied cognitive science. That is, my account proposes that human cognitive systems deal with the control problem by utilizing both centralized and distributed control mechanisms that communicate via a simple message-passing strategy.

In Chapter 6 and Chapter 7, I will analyze recent neuroscientific models of decision-making and control using the theoretical lenses provided by formal decision theory and social choice theory, as well as a legacy of philosophical literature in which the mind is understood using the metaphor of a society. I demonstrate that the human mind can be productively modeled as a society of neural "agents" that collaborate with one another using "democratic" procedures. In particular, I argue that some of the more important democratic procedures are implemented in the basal ganglia, ensuring that a special set of decisions, prior to their execution, are subject to processes that are similar to consensus-seeking and voting. By actively managing and taking advantage of the wisdom-of-the-crowd effect, these democratic processes enhance the intelligence and coherence of the control decisions.

Neurodemocracy not only offers an alternative to classical cognitive science's solution of central systems, but it also goes beyond embodied cognitive science's idea of pandemonium competition. The pandemonium competition model contends that political power struggles between self-interested coalitions of neural agents produce intelligent and coherent control decisions, but does not explicitly specify how this is done. Instead, drawing on formal decision theory and social choice theory, my account stresses that intelligent and coherent control decisions emerge from the democratic cooperation—as opposed to competition—of less intelligent sensorimotor processes.

Drawing upon recent literature on the basal ganglia (Ashby, Turner, & Horvitz, 2010; Wiecki & Frank, 2013), I will also suggest ways in which an empirically testable, mechanistic model that implements these democratic procedures can be developed. Roughly, unlike the classical architecture's central systems, which are constantly privy to all information in the cognitive system and conduct all control decisions, the basal ganglia work like a "central election commission": they delegate mundane control decisions to cortical and other subcortical mechanisms for distributed control. Yet, when novel situations arise, they will engage and determine the appropriate control decisions based on principles of neurodemocracy and simple evaluative information (the votes) received from across the cognitive system. My account will contribute to a coherent scientific image of mind by synthesizing philosophy with the cognitive sciences—in particular, with the emerging field of computational cognitive neuroscience.

5. Structure of the Thesis

This thesis has three parts. Part I is mainly concerned with the control problem faced by massively modular architecture. In Chapter 2, I articulate the key commitments of classical cognitive science's Standard Account, before evaluating its solution to the control problem. Against this background, I then introduce and motivate the massive modularity hypothesis. Chapter 3 identifies an explanatory gap related to the control problem in the massively modular architecture. Having articulated this problem, I provide two arguments that this architecture, owing to its commitment to nativism, lacks the resources to solve the control problem.

Part II revises the society of mind account based on recent cognitive neuroscientific findings for the purpose of providing an up-to-date assessment of embodied cognitive science's progress on the

control problem. I begin Chapter 4 by briefly reviewing Dennett's Pandemonium architecture, showing that it provides an embodied solution to the control problem. Then, I construct the hierarchical embodied cooperative architecture (HECA) by articulating three theses: the Hierarchical Structure Thesis, the Embodied Agent Thesis, and the Cooperative Decisions Thesis. In Chapter 5 I provide empirical support for HECA and furnish it with mechanistic details. I end the chapter by evaluating HECA's progress on the control problem, and I highlight the remaining challenges it faces.

Part III develops my positive account of neurodemocracy. In Chapter 5, I articulate the criteria that a mechanism must meet in order to count as a central controller, and argue that the basal ganglia fulfill them. I am careful to demonstrate, however, that the basal ganglia are significantly different from Fodorian central systems because they use a simple message-passing strategy and delegate control to other distributed controllers. I conclude Chapter 6 by highlighting that while neurodemocracy improves HECA's solution to the control problem, it nonetheless leaves unaddressed some of the challenges of control that arise in novel contexts. In Chapter 7, I conclude the thesis, and address these remaining challenges. Specifically, I provide some empirical support for a more speculative development of my neurodemocracy account based on social choice theory, and explain how this development may help address these remaining challenges posed by the control problem.

Part I

Massive Modularity and the Nativist Information Control Problem

2

Massive Modularity and the Nativist Information Control Problem I: Theoretical Context

1. Introduction

One obstacle to understanding human intelligence is our incomplete knowledge of how the cognitive system solves the control problem: facing open-ended environmental challenges, how do the numerous and diverse components of human cognitive systems coordinate in a manner that ensures the right components are engaged at the right time? Despite its importance, the control problem has been under-examined by cognitive scientists. It seems that, until one attempts to build a cognitive architecture for general intelligence, one can ignore the problem or delegate the task of control to a homunculus (Figure 2.1) (Eliasmith, 2013; Newell, 1980). Hard-wiring all connections between components isn't feasible, and even if it were, we still face a problem: given the characteristic flexibility of human thought, intelligent control mechanisms must determine how to appropriately route information based on the tasks and environments encountered.

There is a debate in philosophy of cognitive science between massive modularity (MM) theorists and their opponents over whether MM models can handle information control (or just "control" henceforth). MM is the thesis that the human mind is composed exclusively of a great many specialized cognitive mechanisms or modules (Barrett & Kurzban, 2006; Carruthers, 2006; D. Sperber, 2007; Tooby & Cosmides, 2016). Fodor maintains that the MM hypothesis is problematic because control cannot be managed exclusively by modules. Specifically, Fodor argues that insurmountable information control problems arise for MM models due to (1) an architectural feature—that individual modules can only receive a limited range of inputs—and (2) an epistemic commitment—that modules implement unreliable heuristics (Fodor, 2000). In response to Fodor's attack, MM theorists maintain that Fodor's characterization of modules is unnecessarily restrictive

and that his standard of human reliability is unrealistic (Barrett, 2005a; Carruthers, 2006; Pinker, 2005; Dan Sperber, 1994). They contend that recent MM models, incorporating self-organization, development, and learning, can handle control satisfactorily and explain human intelligence successfully. Notwithstanding these recent developments, Richard Samuels (2012) argues that MM theorists, in addressing the control problem, have adopted a self-defeating strategy, since they unwittingly incorporate non-modular control structures into the model.

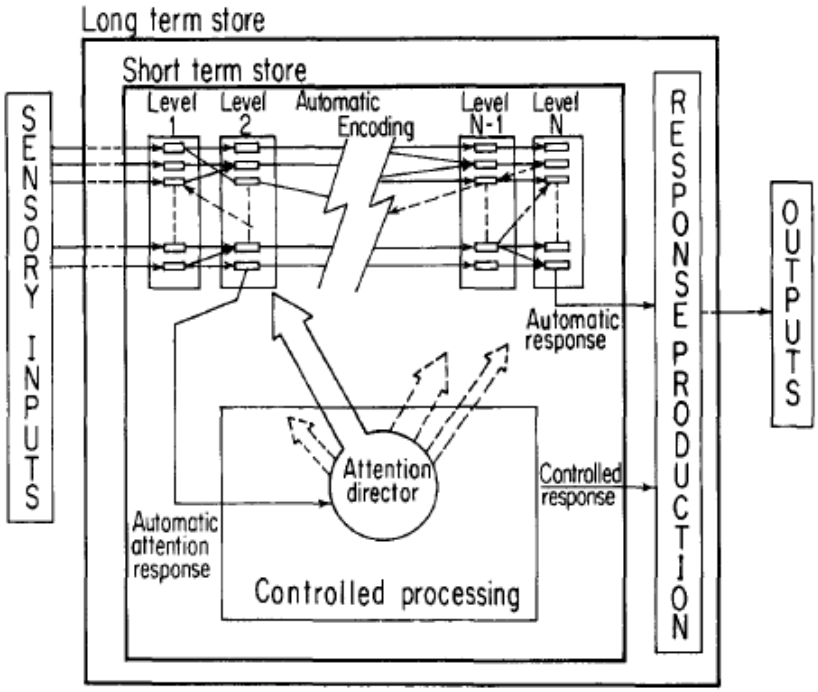

Figure 2.1 The attention director of the controller processing work in this model as a control homunculus. Excerpted from (Shiffrin & Schneider, 1977)

This debate about the necessary features of flexible control mechanisms remains inconclusive. In this chapter and Chapter 3, I will make some headway. My strategy is to argue that MM's real problem when it comes to information control is its commitment to nativism. I argue that the nativism thesis precludes modules from performing control reliably in a wide range of novel contexts where human reason thrives. Hence, in spite of the expanded theoretical resources mentioned above, recent MM models still cannot handle control flexibly enough to explain human intelligence. Placing the debate within the context of control will also enable me to extract lessons for mechanistic accounts of the neural control structure.

I will start (Section 2) by refining the concept of behavioral flexibility, which is a key feature of human intelligence. I reveal that MM only seems to successfully explain human intelligence, and that this illusion of explanatory power owes to widespread confusion over what behavioral flexibility really is. Section 3 will review the *Standard Account* of cognitive architecture to illuminate the theoretical context against which MM reacts. I will also briefly discuss how the Standard Account addresses the control problem. In Section 4, I will review MM's architectural commitment in more detail. Doing so will help make apparent, in Chapter 3, that proponents of MM have the theoretical resources to circumvent important existing challenges to their account. Section 5 will highlight MM's epistemic commitment and relate it to its architectural one. After clarifying the explanandum (behavioral flexibility) and explanans (MM) in this chapter, I will argue in Chapter 3 that there exists an explanatory gap that MM has no theoretical resources to fill due to its nativist commitment.

2. What Behavioral Flexibility Really Is

Behavioral flexibility is an important aspect of human intelligence. Indeed, cognitive architectures, in explaining human intelligence, often take it as the key explanandum. As we will see, a misunderstanding of behavioral flexibility lies behind many problematic arguments advanced by both MM theorists and their opponents. In this section, I will provide a more adequate analysis of this concept.

Let us begin with the following characterization:

> **Intuitive Conception of Behavioral Flexibility**: Humans have the competences to perform tasks satisficingly in a context-appropriate fashion, in a wide range of novel and significantly different contexts.[17]

This characterization captures our intuition about human intelligence. For example, we can communicate appropriately in a diverse range of social situations; we learn to perform many arbitrary and novel tasks, be it chess or air-traffic control, and do so with high levels of expertise. MM theorists seem to agree. Clark H. Barrett and Robert Kurzban write, for instance, "Humans unquestionably face and solve challenges their ancestors never faced" (2006, p. 635). To be clear about what the intuitive conception of behavioral flexibility asserts, let us consider its details.

The intuitive conception claims that humans have the competences to perform tasks satisficingly in a context-appropriate fashion. To perform a task satisficingly is to produce good enough results often enough according to the standard imposed by the relevant context, but not necessarily results that are optimal or even optimal under constraints.[18] "Context" here refers to the information structure of a task in a particular environment, e.g., task-relevant functional properties, the pattern of covariations between them, and their detectable cues in the environment.[19] It is important to note that the same type of task, however one may want to individuate tasks, may have different information structures in different environments, and therefore constitute different contexts. Moreover, because information structures determine the competences required to perform the task, the same type of task in different environments may also demand different competences. For example, the task of driving can have slightly different information structures depending on whether one is driving a manual or automatic car; as a result, the required competences are different too. This is a crucial fact often ignored by MM theorists.

Novel contexts are non-adaptive contexts, where adaptive contexts refer to the information structures of *adaptive problems* in the *environment of evolutionary adaptedness* (EEA) (Tooby & Cosmides, 1995a). According to MM theorists, adaptive problems are recurring problems whose

---

[17] Roughly, competences refer to the core knowledge required to perform relevant tasks, such as Chomsky's Universal Grammar for language (Chomsky, 1965, 1975, 2005). Competences can be either innate or acquired.

[18] A performance is optimal if its result maximizes some value or criterion, such as expected utilities. A performance is optimal under constraints if it is the best one can do, taking into consideration the tradeoff with cognitive or environmental resources (Gigerenzer, 2006), and genetic or developmental constraints.

[19] My notion of information structure is similar to the notion of environment structure when one studies the ecological rationality of heuristics (Todd & Gigerenzer, 2012). However, my notion is broader in the sense that it not only includes the structures of external and internal environments, but also that of the task. For example, the information structure of cheater detection may depend on the structures of the external environment, the internal cognitive mechanisms of the reasoner, and the task itself.

solutions promote fitness directly or indirectly (e.g., mating, foraging, navigation, etc.). The EEA refers to the ancestral environments that shaped the design of an adaptation in pre-human and early human history.[20] Given these remarks, we should note that some contexts in the modern world may share the same information structures as adaptive ones. When this is the case, such modern contexts are not novel.

The intuitive conception of behavioral flexibility can be distinguished from other conceptions of behavioral flexibility. In particular, we can differentiate the intuitive conception from the following two notions (see Figure 2.2): [21]

> **Optimistic Conception of Behavioral Flexibility**: We have the competences to perform tasks satisficingly in a context-appropriate fashion, *in all possible contexts* (given perhaps unlimited resources and time).
>
> **Pessimistic Conception of Behavioral Flexibility**: We have the competences to perform tasks satisficingly in a context-appropriate fashion, *in a small range of novel contexts that are similar to adaptive ones.*

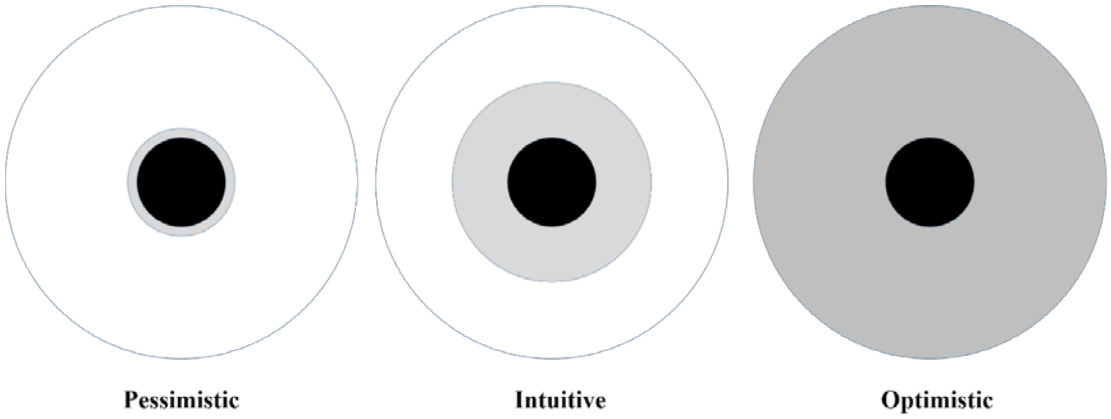

Figure 2.2 Different conceptions of behavioral flexibility. The black circle represents the adaptive contexts in which humans exhibit satisficing performance. The gray area represents the novel contexts in which human beings exhibit satisficing performance. The white area represents all other possible contexts. For the optimistic conception, the gray and white areas overlap completely.

The optimistic conception is taken to be the hallmark of human intelligence by some philosophers (Descartes, *Discourse on Method*, 5; Horgan & Tienson, 1996) and cognitive scientists (J. R Anderson & Lebiere, 2003; Newell, 1994) This position seems, however, unduly strong and, in any case, I do not need to endorse it in order to make my case against MM. As such, my arguments do not assume the optimistic conception of human intelligence. On the other hand, the pessimistic conception is problematic because it claims that the diversity in human problem-solving is merely apparent. So neither do I assume the pessimistic conception. (For readers who think I dismiss the

---

[20] The Pleistocene is often regarded as the EEA for the modern human brain (Richerson & Boyd, 2012). I remain neutral about any methodological issues concerning reliable identification of adaptive problems (Atkinson & Wheeler, 2004; Richardson, 2010; Sterelny & Griffiths, 1999) and my arguments do not depend on such issues.

[21] it is worth noting that these different conceptions, in particular the intuitive and pessimistic ones, are not quantitatively defined positions. Instead, they are qualitatively significant regions on a spectral scale of flexibility. Because the evidence I appeal to is similarly coarse-grained and qualitatively significant, these characterizations of flexibility are sufficient for the purpose of my arguments.

pessimistic conception too hastily, I will address two relevant objections in Chapter 3) Taking a middle ground between these two positions, I will assume the intuitive conception and refer to it simply as "flexibility" henceforth.[22] With a clarified conception of the key explanandum, I will now turn to one of the major explanans, the Standard Account of cognitive architecture, in order to provide a better theoretical context for discussing MM.

3. The Standard Account

In this section, I will focus on clarifying the Standard Account of cognitive architecture in order to provide a clear view of a major opponent of MM. The origins of the Standard Account can be traced to the early development of Artificial Intelligence and Robotics. Its core commitment is to an epistemological theory of human reasoning—ideal non-demonstrative reasoning competence. I shall refer to this competence as the "epistemic commitment" of the Standard Account. To implement this epistemological theory, central systems—computational mechanisms with a distinctive set of architectural features—are proposed. I shall refer to the central system as the "architectural commitment" of the Standard Account. This approach can be found more or less explicitly in the work of many cognitive scientists. For example, the epistemic commitment is spelled out in Fodor's earlier work (Fodor, 1975), while the architectural commitment is articulated in Miller's work on planning, Newell and Simon's research on problem-solving, and Anderson's ACT-R cognitive architecture (J. R. Anderson, 1983; Miller, Galanter, & Pribram, 1960; Newell & Simon, 1972).[23]

In the following section, I will begin by articulating the Standard Account's epistemic commitment: Ideal non-demonstrative reasoning competence. I will then discuss central systems, its key architectural commitment. By the end of Section 3 it will be obvious that if the Standard Account were true, it could handle the control problem and explain human flexibility. As we shall see, however, the Standard Account faces two fatal challenges: the problem of performance and the problem of competence. I end this section by reviewing these two problems, against which we can best understand MM's theoretical motivation.

3.1. The Standard Account's Epistemic Commitment

In this section, I will discuss the Standard Account's epistemic commitment: competence for ideal non-demonstrative reasoning.

Non-demonstrative reasoning is involved in many typical cognitive tasks, including perception, learning, and planning. It is through the exercise of non-demonstrative reasoning that a cognitive agent acquires a new belief or intention that they did not have before (or abandons a belief or intention to which they were originally committed). The belief and intention acquired or abandoned through this process are typically "logically independent" from the agent's prior cognitive or conative commitments—hence the term "non-demonstrative" reasoning. Take, as an example of theoretical reasoning, the fixation of perceptual belief. This process starts with the

---

[22] I will grant MM representational flexibility (Machery, 2008; Samuels, 2012) and argue that despite this, it nevertheless fails to explain behavioral flexibility.

[23] I am following Samuels (2010) in calling this prominent account, the 'Standard Account'. We should note that these two features of the Standard Account need not always go together. Some cognitive architectures implement Standard Account's architectural features without implementing its epistemic theory (John R. Anderson, 1983; Miller, Galanter, & Pribram, 1960; Newell & Simon, 1972).

registration of sensory information, from which beliefs about the sensory information's distal causes are inferred non-demonstratively. As Fodor describes it:

> Since the relation between a distal object and the sensations it causes a perceiver to have are (typically; maybe always) contingent in both directions, the inferences that mediate the fixation of perceptual beliefs are (typically, maybe always) contingent too. (Fodor, 2008, p. 113)[24]

An example of practical reasoning is action planning. According to the Standard Account, action planning works in roughly the following way: First, one generates a list of candidate actions. For each candidate action, a list of outcomes and the probability that the candidate action will lead to each of them are calculated. Then, the value of each outcome is assessed by calculating the utilities of the states of affairs involved. After this, the "expected value" of a candidate action is determined by multiplying the value of each outcome by the probability the action will lead to it and summing the numbers up. Finally, the candidate action with the highest expected value is adopted for execution. Because the relations between a candidate action and its resulting "hypothetical worlds" are also "contingent" in both directions, the process of action planning in general (the process of working out the potential outcomes of an action in particular) is a species of non-demonstrative reasoning.

An ideal non-demonstrative reasoning process is one that produces optimal solutions in all possible contexts. It should be obvious that our capacity for ideal non-demonstrative reasoning can explain our behavioral flexibility.[25]

The capacity for ideal non-demonstrative reasoning can be analyzed into two components: reasoning competence and performance. The epistemic commitment of the Standard Account is the reasoning competence, an internally represented theory of ideal non-demonstrative reasoning.

The competence/performance distinction was first introduced to cognitive science by Chomsky as an explanatory strategy in understanding "linguistic intuitions," the primary data of linguistics (Chomsky, 1965, 1975, 2005). Linguistic intuitions are unreflective judgments about grammaticality and other linguistic properties of sentences. In order to explain these intuitions, as well as how speech is produced and comprehended more generally, Chomsky proposed a hypothesis that would become one of the most influential theories in cognitive science. According to this hypothesis, a language speaker possesses "linguistic competence," an implicit knowledge of language. Their linguistic competence is composed of the internally represented grammar—a set of generative rules and principles—of the particular language they speak. When a language user makes an intuitive linguistic judgment, information encoded in their linguistic competence is implicitly accessed and relied upon. However, linguistic competence cannot produce linguistic intuitions by itself. In a complex process of intuition production, it needs to interact with a variety of other cognitive mechanisms, including those responsible for perception, attention, motivation, memory, etc. These other cognitive mechanisms are referred to as the "linguistic performance mechanisms." Performance mechanisms, together with the linguistic intuitions and the sentences the speaker actually produces and comprehends, constitute "linguistic performance." In some situations, activity

---

[24] That is, "contingent" as "not logically necessitated from our epistemic point of view".

[25] In fact, the capacity of ideal non-demonstrative reasoning can explain not just the intuitive conception of flexibility; it can explain the optimistic conception of flexibility—the competences to perform tasks satisficingly in a context-appropriate fashion, *in all possible contexts*.

in any of the performance mechanisms may lead a language-user to make a false linguistic judgment that does not reflect their linguistic competence. In such cases, we say that this person has made a "performance error."

Because there are some close analogies between the phenomena studied in linguistics and those studied in the cognitive science of reasoning, it is plausible to apply the competence/performance distinction in the explanation of our capacity for reasoning (Samuels, Stich, & Tremoulet, 1999). Competence for ideal non-demonstrative reasoning consists in an internally represented, integrated set of rules and principles of reasoning that are accessed and relied upon when cognizers make relevant inferences and judgments. Moreover, this integrated set of rules is optimal, in the sense that applying it correctly to any non-demonstrative inference task in any context will result in optimal solutions—provided, that is, that all relevant hypothesis/options and true beliefs (and perhaps also unlimited time and resources) are made available.

Exactly what this set of rules and principles looks like is still a topic of debate among cognitive scientists.[26] However, for the purpose of illustration, it is possible that reasoning competence consists partly in some or other variety of the Bayesian theory of rational belief acceptance (Kaplan, 1981). According to this theory, (1) a rational enquirer should assign probabilities to each of her hypotheses in light of the probabilities she has already ascribed to her existing background beliefs; (2) She also needs to update these probabilities as new evidence comes in, in ways that reflect her assessment of both the independent probabilities of the evidence and its probabilities given the hypotheses are true. As a result, the probabilities an enquirer assigns to any hypotheses, at any given time, will depend on the probability assignment made to her background beliefs at that time. The above two steps will be governed by a set of optimally reliable rules so that the subjective probabilities assigned to the background beliefs and hypotheses will achieve an optimal match with the objective probabilities of events in the world.

As in the case of linguistic competence, competence for ideal non-demonstrative reasoning cannot produce an inference or judgment by itself. Rather, it needs to interact with other performance mechanisms. Many activities in the performance mechanisms can go wrong: relevant hypotheses/options may not be generated, one may not possess relevant true beliefs in the memory, or the agent may simply lack adequate motivational or attentional resources. When performance errors happen, less than ideal reasoning results. Nevertheless, the Standard Account argues that we must possess ideal non-demonstrative reasoning competence because we do reason in accordance to the ideal theory sometimes (and do not do so accidentally).

In this section, I have introduced the Standard Account's epistemic commitment: ideal non-demonstrative reasoning competence. However, this competence needs to be implemented by cognitive mechanisms with suitable features in order to function properly. In the next section, I will look into the architectural features required to implement the Standard Account's epistemic commitment.

3.2. The Standard Account's Architectural Commitment

In this section, I will examine the Standard Account's architectural commitment to central systems. Both the Standard Account and MM architecture assume the existence of modular computational

[26] Actually, for all we know we may not even have such ideal reasoning competence, as we will see soon in the following discussions.

mechanisms for perceptual processing, as well as action production and control. In contrast to MM approaches, however, it assumes the existence of non-modular computational mechanisms, central systems, for theoretical and practical reasoning. Importantly, central systems possess a distinct set of architectural features that make them a particularly suitable type of computational mechanism for implementing the competence for ideal non-demonstrative reasoning.[27] Two of these features are domain-generality and informational non-encapsulation (Fodor, 1975, 1983).

Central systems are domain-general in the sense that they can receive inputs and perform tasks from a wide range of domains (Fodor, 1975, 1983). Because ideal non-demonstrative reasoning competence is optimal in all contexts, it is natural that the central systems which implement it should be given the responsibility to handle tasks from a wide range of domains.[28] Also, because the central systems handle many different domains of tasks, it is necessary that they receive inputs from a wide range of domains.

Central systems are informationally unencapsulated, in the sense that they can make use of all relevant and available information in the cognitive system. As our reasoning competence can reliably produce optimal results only when (almost) all relevant information is made available to it, central systems should have the architecture to take all relevant information into account.[29] As we will see in the next section, the architectural commitment of the Standard Account stands in stark contrast to that of MM, according to which all computational mechanisms are domain-specific and informationally encapsulated. The Standard Account's architectural commitment is conceptually connected to its epistemic commitment, and, as I will explore further in section 5, the same can be said of MM. As a result, the tension between Standard Account and MM is an architectural one as much as an epistemic one.

One comment: The Standard Account, as we've seen, is jointly characterized by its epistemic and architectural commitments—competence for ideal non-demonstrative reasoning and central systems. However, some variations of the Standard Account have been advanced that step back from the epistemic commitment while hanging on to the architectural one—an example of this is the Library Model of Cognition, which I will discuss at the end of Chapter 3 (Samuels, 2005).[30] Although the exact extent to which they back down on the epistemic commitment is unclear, it is fair to say that the further they withdraw from it, the less clear it is that they have the resources to explain behavioral flexibility. After all, it is competence for ideal non-demonstrative reasoning that lies at the center of the Standard Account's explanation of flexibility.

---

[27] However, the central system's architecture is neither sufficient, and perhaps not necessary, for implementing the competence for ideal non-demonstrative reasoning. It is perfectly plausible to have a central system implementing exclusively some kind of non-ideal non-demonstrative reasoning competence. Also, while it is clear that the central system is *one* way of implementing the competence of ideal non-demonstrative reasoning, it may not be the *only* way. There are certainly other proposals on the table, e.g., quantum computation, analog computation, etc. At this stage, it is not clear whether we should take these proposals seriously. In any case, the main argument of this chapter is not affected.

[28] Moreover, it is also natural that these different domains include ones beyond the adaptive problems human beings faced in Pleistocene.

[29] That is, even if performance errors/factors prevent such an ideal to be achieved sometimes or even most of the time.

[30] Samuels suggests a variant of the Standard Account, according to which central systems do not implement competence for ideal non-demonstrative reasoning, do not run exhaustive search for all relevant information, and have limited sensibility to the global properties of cognition. (Samuels, 2005, pp. 120–1)

In short, the Standard Account can be jointly characterized by its epistemic and architectural commitments: the competence for ideal non-demonstrative reasoning and central systems.

3.3. The Standard Account's Solution to the Control Problem

The Standard Account's architecture, with central systems that implement ideal non-demonstrative reasoning competence, promises the perfect solution for the control problem and has the resources to explain behavioral flexibility easily. However, as we will see, the central systems are too perfect to be true of human cognitive systems.

With regard to the problem of intelligence (discussed in Section 2.2 of Chapter 1), the Standard Account can address it easily. The problem of intelligence is the challenge for neural mechanisms to work together to produce intelligent control decisions at various loci in a cognitive system. The Standard Account addresses this issue by concentrating all difficult control decisions in the middle, that is, by assigning them to central systems. The central systems can deal with all of these difficult decisions because they are attributed ideal non-demonstrative reasoning competence and the capacity to access all available information in the cognitive system. These two features of central systems are jointly sufficient (barring performance errors) for optimal practical and theoretical decision-making.

Moreover, the Standard Account can handle the problem of coherence (also discussed in Section 2.2 of Chapter 1) which concerns how coherent control decisions can be generated by an architecture that is composed of highly fragmented and distributed neural mechanisms. The Standard Account deals with this issue by (1) granting central systems the ability to issue all important control decisions and set the ultimate goals of the cognitive system, and (2) reducing modules to central systems' "slaves," which can only work out the means to the goals set by the central systems. As the central systems maintain (more or less) unified models of the world and its reward structures, the cognitive system under their dictatorship will perform (more or less) coherent behaviors.

In short, if the Standard Account were true, it could have easily addressed the control problem and explained behavioral flexibility.[31] However, the Standard Account suffers from two fatal problems, both of which stem from its epistemic commitment: one is the problem of performance; the other is the problem of competence.

3.4. The Problem of Performance

One of the most prominent challenges to the Standard Account is based on the consideration of tractability. Although there are different versions of this argument with significant variance in details, they all share the same basic structure. They argue that the capacity for ideal non-demonstrative reasoning cannot be implemented in classical computational architecture within realistic constraints of tractable computation. Because classical computational architecture is

[31] With regard to the problem of architecture, the central systems' architectural features can address it adequately as well. The architectural issue (discussed in Section 2.2 of Chapter 1) refers to the challenge of how connections between cognitive mechanisms can be set up adequately and economically so that the cognitive mechanisms can be recruited flexibly to solve a wide range of novel tasks. Because all perceptual modules are (directly or indirectly) connected to the central systems, which then (directly and indirectly) connect to all motor modules, the central systems have the necessary connective structure to communicate commands and transfer information to all modules in order to selectively recruit them for open-ended problem-solving.

necessary for naturalistically implementing ideal non-demonstrative competence (Fodor, 2000), the Standard Account cannot be true of the human cognitive system (in so far as the human cognitive system is realized naturalistically). Because the problem lies in the implementation of competence—that is, in the relevant performance mechanisms—I call it "the problem of performance." In the following, I will briefly discuss a version of the argument, known as *the problem of relevance*.

It starts with a quite uncontroversial premise:

> (1) An ideal non-demonstrative reasoning process needs to consider all relevant information to produce optimal solutions reliably.

For example, in hypothesis confirmation, the reasoning process needs to attend to all available and relevant information that would affect the estimated subjective probability of the hypothesis in order to reliably produce an optimal result. However, the crucial issue concerns how the cognitive system decides the relevance of information. This issue is particularly problematic because:

> (2) Due to the nature of non-demonstrative reasoning, (almost) any piece of information, given appropriate background beliefs, can be relevant to the task at hand.

That is, the typical non-demonstrative inference is isotropic (Fodor, 1983, 2000). For example, whether it will be a full moon next weekend seems to have little to do with whether or not I will embark on a peaceful vacation to a Pacific island then. Yet, suppose that I happen to believe that there is a beach party held on every full moon night, one that will destroy the tranquility I hope to enjoy during my trip. It thus seems that my belief that it is going to be a full moon next weekend is relevant to how likely it is that I will have a peaceful vacation. *Mutatis mutandis* for other cases. Now, because non-demonstrative inference is isotropic, there is no way of deciding *a priori* which information is relevant, and "any substantive criterion of relevance is at risk of eliminating something that is in fact germane; that is, something that if it *were* attended to *would* affect the estimated subjective probability of the belief" (Fodor, 2008, p. 116). Actually, relevance is a context-dependent property: whether or not a piece of information is relevant depends on epistemic context, i.e., the background beliefs the information is embedded in. "As things now stand, Classical architectures know of no reliable way to recognize such properties short of exhaustive searches of the background of epistemic commitments" (Fodor, 2000, p. 38). As a result,

> (3) The only way for the reasoning process to take all relevant information into account is to actually assess (almost) all of the available information (Dietrich & Fields, 1995).

That is: "Reliable abduction may require, in the limit, that the whole background of epistemic commitments be somehow brought to bear on planning and belief fixation" (Fodor, 2000, p. 37). This would not be a problem if the size of a person's background beliefs were small, but it is not. Human beings typically possess a richness of epistemic commitments. Consequently,

> (4) Bringing the entirety of one's epistemic commitments to bear would require more than tractable computation in a classical computational architecture.

While tractability was originally a technical term in computer science, the notion of tractable computation involved in cognitive science is different.[32] A *tractable computational process* is one that does not require more time and resources (memory, information, computation power, etc.) than what a normal human being can be expected to possess for the completion of a task. Although this notion of tractability cannot be made more exact,[33] it is nonetheless clear that bringing one's entire background beliefs to bear on a non-demonstrative reasoning task would typically be intractable. For example, consistency checking for any candidate new belief is traditionally assumed to be part of any ideal non-demonstrative reasoning task.[34] Yet consistency checking is intractable if attempted on an exhaustive basis even for a set of just 138 beliefs (Cherniak, 1990, p. 93). Consider checking the consistency of beliefs using a truth table. Even if each line can be examined in the time that it takes a photon of light to travel the diameter of a proton, it would still take more than 20 billion years to complete 2^{138} lines. As Fodor puts it,

> The totality of one's epistemic commitments is *vastly* too large a space to have to search if all one's trying to do is figure out whether, since there are clouds, it would be wise to carry an umbrella. Indeed, the totality of one's epistemic commitments is vastly too large a space to have to search *whatever* it is that one is trying to figure out. (Fodor, 2000, p. 31)

As a result, it is reasonable to conclude that,

> (C) Human reasoning cannot be the classical computational implementation of the capacity for ideal non-demonstrative reasoning.

In short, the problem of performance shows that the Standard Account cannot be a plausible explanation of human flexibility. This is because ideal non-demonstrative reasoning competence cannot be implemented in human cognitive systems within reasonable biological constraints.[35]

---

[32] The concept of tractability in computer science is used in various non-equivalent ways. According to one characterization of tractability from the computational complexity literature, an algorithm for solving a given problem is tractable if, in the worst case, the number of steps needed for completing it is a polynomal (or better) function of the input size. An algorithm is intractable if it is not tractable. However, such a characterization is not suitable for our purpose as cognitive scientists, according to Samuels (2010, p. 282) for two reasons. First, pretty much any interesting algorithm that models human cognitive process is intractable according to this criterion. So, intractability does nothing more than characterize the phenomena of interests to us. Second, an intractable algorithm normally takes much less steps to complete than its worst-case scenario. Thus, if some performance limitation prevents the algorithm from being used in the worst case, being intractable would not pose a serious problem.

[33] Although common-sense reflection can give us some sense of the real-time constraint of our cognition, we still lack sufficient information about memory capacity, the speed of processing, and other important parameters of our cognitive system to make the concept of tractability exact.

[34] Even if consistency checking turned out not to be part of the ideal reasoning process, it would still be clear that it involved much simpler algorithm than any ideal non-demonstrative reasoning task did, say, Bayesian confirmation.

[35] One way to respond to this conclusion is giving up the naturalistic approach to understanding human intelligence. According to Fodor (1983, 2000, 2008), because (1) we obviously have the capacity for ideal non-demonstrative reasoning, and (2) classical computational approach is our only hope for understanding its implementation naturalistically, consequently, human reasoning process as well as the operation of central systems will remain a mystery for us. However, there is another way to respond to this conclusion. According to a more optimistic stance, the conclusion does not cause such epistemic crisis. It is because the real take-home message of this argument is that perhaps, we do not possess such ideal reasoning capacity after all. In the following section, we consider the argument for this more optimistic stance.

3.5. The Problem of Competence

In addition to the problem of performance, according to which human cognitive systems, given the biological constraints, cannot implement ideal reasoning competence, there is an alternative challenge faced by the Standard Account. Even if human cognitive systems *could* implement ideal non-demonstrative reasoning competence, it is an empirical fact that they do not (Samuels et al., 1999). Because this objection against the Standard Account denies that human cognitive systems implement ideal non-demonstrative competence, I shall call it the *problem of competence*.

The argument starts with an empirical observation that has been reported by psychologists for the past four decades:

> (1) Human reasoning performance deviates systematically from the dictates of the normative theory that supposedly constitutes ideal reasoning competence under ordinary circumstances.

Psychologists find that people are extremely bad at many types of reasoning tasks. For example, experiments involving the Wason selection task suggest that people have serious problems identifying the situations under which a conditional statement is true (Evans & Over, 1996). Human subjects, when reasoning about probabilities, also commit numerous reasoning fallacies, such as the conjunction fallacy and base-rate neglect (Tversky & Kahneman, 1974). For example, according to the Bayesian account of confirmation, the subjective probability of a hypothesis is determined in part by the prior probability of the hypothesis. However, Tversky and Kahneman show in a series of experiments that the base-rate information, which would inform the prior probability of a hypothesis, is entirely ignored by subjects in cases where some specific, but completely worthless, information is provided to them. "People's grasp of the relevance of base-rate information must be very weak if they could be distracted from using it by exposure to useless target case information" (Nisbett & Ross, 1980, pp. 145-6).

These mistakes are made even when factors such as fatigue and strong emotions are not present, and the errors made in the experiment setting are very different from those made when subjects' memory or attentional resources are exhausted, or when they are under the influence of drugs. According to one pessimistic interpretation, these experimental results have "bleak implications" for our reasoning capacity (Samuels, 1998b). Specifically, they support the hypothesis that:

> (2) Because these deviant reasoning processes are not caused by performance errors, they must be the result of competence errors.[36]

The subjects are in fact reasoning in accordance with their reasoning competence when they make mistakes, but their reasoning competence cannot handle a wide range of reasoning tasks. They do not use the right principles because they have no access to them, because they do not form part of their reasoning competence. That is,

> (C) Human subjects do not possess ideal non-demonstrative reasoning competence.

---

[36] Strictly speaking, premise (2) does not follow from (1), because memory, attention, emotion, etc. only constitute a small part of the performance factors. The deviant reasoning performance may still result from other types of performance errors.

In sum, the problem of competence shows that the Standard Account cannot be the true picture of the human mind because empirical evidence strongly suggests that the human cognitive system does not implement ideal non-demonstrative reasoning competence.

Both the problem of competence and the problem of performance show that the Standard Account offers both a flawed picture of the human cognitive system and a problematic explanation of human flexibility. We cannot be the ideal reasoners we take ourselves to be. Instead, what we possess has to be a less-than-ideal reasoning competence. As a result, studies into the nature of human mind should abandon the Standard Account in favor of one that is grounded in a realistic consideration of our bounded rationality. It is against this background that we can best understand the research program of MM and its epistemic and architectural commitments.

4. The Massive Modularity Hypothesis: Architectural Commitment

This section will spell out MM selectively by focusing, to begin with, on its architectural commitment. I aim to provide MM as many theoretical resources as possible by avoiding non-essential architectural features—this will allow MM to sidestep some existing criticisms. In the next section, I will turn to MM's epistemic commitment, which both MM theorists and their opponents often fail to fully appreciate—this will prepare us to understand MM's real limitation to explaining flexibility in the next chapter.

MM is a thesis endorsed by many evolutionary psychologists who are committed to nativism (Barkow, Cosmides, & Tooby, 1995; Barrett & Kurzban, 2006; Buss, 2005; Pinker, 1999; Dan Sperber, 1994):[37]

> **Massive Modularity Hypothesis (MM)**: The human mind is composed exclusively of a massive set of Darwinian modules.

Let me clarify MM in more detail. MM is, according to Samuels, the conjunction of the following three theses (1998b):

1) **Massive modularity:** the human mind contains massive numbers of Darwinian modules

2) **Central modularity:** not only are there modules for peripheral processes (e.g., input/output modules for perception and action), there are also modules for central processes (i.e., modules for reasoning and belief-fixation), such as social reasoning (Cosmides & Tooby, 1992), biological categorization (S. Pinker, 1995), and theory of mind (Baron-Cohen, 1995), etc.

3) **Exclusive modularity:** the mind is made up exclusively of Darwinian modules. It contains no non-modular computational mechanisms (e.g., central system)

Weaker versions of massive modularity often reject the Exclusive Modularity thesis, and some reject both the Exclusive Modularity and Central Modularity theses. Although they are empirically

[37] I will spell out different versions of nativism in Chapter 3.

viable positions, following Samuels (1998b), I will not be concerned with them in this chapter.[38] In the following discussion, we will only be concerned with the strong version of massive modularity (MM).

4.1. What Darwinian Modules Are Not

Before I elaborate on the features of Darwinian modules, let me say a few things about what Darwinian modules are not. Darwinian modules, as a type of *computational mechanism,*[39] need to be clearly distinguished from all types of *informational modules*, that is, systems of mentally represented information (Samuels et al., 1999). The paradigm example of an informational module can be traced back to Chomsky's work in linguistics. A Chomskian module is an informational module that is (1) truth-evaluable: it makes sense to ask of these mental representations whether they are true or false; (2) innate: these mental representations are not learned; (3) domain-specific: the module is dedicated to solving problems in a restricted domain, and its operations are (4) inaccessible to consciousness (Samuels et al., 1999). Chomsky claims that our linguistic competence consists in an innate and internally represented grammar; "universal grammar" is the quintessential example of a Chomskian module. MM theorists, in characterizing modules as "autonomous mental mechanisms" and "functionally dedicated computers" (Tooby & Cosmides, 2000, p. 1189), indicate that their conception of a module is a computational, as opposed to informational one.

We also need to be careful not to treat all types of computational modules in the literature as *Darwinian* modules. Specifically, we should separate Darwinian modules from two other popular conceptions of computational modules: minimal modules and Fodorian modules.

A minimal module is simply a functionally characterized computational mechanism that has a specific function (for example, face recognition) to perform in the overall cognitive system. According to this minimal construal, "…anything that would have its proprietary box in a psychologist's information flow diagram—thereby counts as a module" (Fodor, 2000, p. 56). It is likely that most mainstream cognitive scientists, with the exception of some connectionists and dynamicists, believe that the mind consists exclusively of modules of this kind (Fodor, 2000, p. 56). While a Darwinian module is of course a type of minimal module, the notion of minimal module is too thin to capture all its features.

---

[38] Following Samuels (2006), I will not discuss the weaker versions of MM for three reasons: First, MM clearly is what prominent evolutionary psychologists have in mind when they talk about their preferred view of mental architecture. As Steven Pinker puts it, the human mind is "not a general-purpose computer but a collection of instincts adapted for solving evolutionarily significant problems—the mind as a Swiss Army knife" (S. Pinker, 1995). Second, arguments for evolutionary psychology do not support weaker versions of massive modularity, as they usually argue that evolution cannot produce non-modular mental structures (Samuels, 2006, p. 42). Finally, while MM is a distinct theoretical position, weaker versions of massive modularity are not. They are compatible with some versions of the Standard Account discussed in the last section.

[39] Clearly, MM theorists adopt a computational approach to psychology that is dominant in contemporary cognitive science. According to this approach, the mind is a computational machine and mental processes are computational processes. As Barkow, Cosmides, and Tooby maintain:

> [The brain is] a computer made out of organic compounds rather than silicon chips. The brain takes sensorily derived information from the environment as input, performs complex transformations on that information, and produces either data structures (representations) or behavior as output. (Barkow, Cosmides, & Tooby, 1995, p. 8)

On the other hand, a Fodorian module, as characterized by Fodor in his monumental work, *The Modularity of Mind,* is a much more demanding notion than a Darwinian module (Fodor, 1983). A Fodorian module is a minimal module that possesses the following features to an interesting degree:

i. **Domain-specificity:** it only takes a narrow domain of representations as input, and functions to solve a narrow domain of tasks.
ii. **Informational encapsulation**: besides its input, a module only has access to information in its own proprietary database.
iii. **Mandatory operation**: one cannot control (directly) the operation of a module.
iv. **Fast speed**: its operation speed is fast compared to a non-modular computational mechanism.
v. **Shallow output:** its output is a preliminary characterization of its input.
vi. **Limited accessibility**: other systems have limited access to a module's intermediate processes and representations.
vii. **Localization**: the neural implementation of a module is highly circumscribed and dedicated to the realization of the module and that module only.
viii. **Characteristic breakdown pattern:** because modules are implemented in localized and dedicated neural circuits, they can be selectively damaged while keeping the function of other modules relatively intact.
ix. **Innateness:** the modules possess the property of "develop[ing] according to specific, endogenously determined patterns under the impact of environmental releasers" (Fodor, 1983, p. 100).

Notably, not even Fodor adopts this demanding characterization of modules in his more recent work (Fodor, 2000).

4.2. Darwinian Modules

Having articulated what Darwinian modules are not, I will now turn to their main features:

> Darwinian modules are domain-specific and informationally-encapsulated computational mechanisms; they are either innate or the products of learning.

In the following, I will discuss the key concepts of "domain-specificity," "information encapsulation," and "innateness" that are involved in defining Darwinian modules (henceforth, simply "modules").

Domain-Specificity

Domain-specificity is a proper-function concept for MM theorists (Barrett & Kurzban, 2006; Buss, 2016; Cosmides & Tooby, 2013). Roughly, the proper function of a trait or structure (e.g., a module) is X if the trait or structure exists and is maintained in a given population by natural selection (or other selection processes, such as biological development and learning) in virtue of performing X (Allen, 2009). A cognitive mechanism (e.g., a cheater-detection module) is domain-specific if it is designed (by evolution) to specialize in a limited range of tasks (its task-domain).

In other words, domain-specificity is a kind of functional specialization. Many evolutionary psychologists understand domain and domain-specificity in this way when they express the view that "our cognitive architecture resembles a confederation of hundreds or thousands of functionally dedicated computers (often called modules) designed to solve adaptive problems endemic to our hunter-gatherer ancestor " (Tooby & Cosmides, 1995a, p. xiii). Relatedly, MM theorists believe

that (most) Darwinian modules are adaptations that have been shaped by natural selection over time to solve specific adaptive problems. As a result, modules come to have impressive functional designs that allow them to solve adaptive problems, at least satisficingly.[40]

A clarification: there is actually an additional distinctive sense of domain that needs to be distinguished from the task-domain. Domain can also be understood as input-domain, which refers to the type of representations a given module can take as input, e.g., visual, auditory representations, or more narrowly, phonetic or face representations. Domain-specificity in this sense is a restriction on the representations a computational mechanism can take as input. A mechanism is domain-specific, under this conception, if it only takes a highly restricted range of representations as input. We should note that in his earlier work, Fodor understands "domain" in both these senses of task and input (Fodor, 1983).[41]

The two notions of domain-specificity are certainly connected. Restricting a module's input domain is one way of narrowing down its task domain. Also, paradigm examples of domain-specific mechanisms, such as those responsible for low-level visual perception and face recognition, seem to count as domain-specific in both senses of the term. However, as my present goal is to characterize the understanding of modules that is essential to MM, I will adopt the task-domain notion of domain-specificity throughout this chapter and Chapter 3.

Informational Encapsulation

The information encapsulation of a Darwinian module is, at its core, a restriction on the amount of information it can use when performing tasks. A mechanism is encapsulated if its information processing cannot be influenced by most of the task-relevant information available in the mind *during the process of a particular task* (Barrett, 2005a; Carruthers, 2006).[42] That is, a Darwinian module is *frugal*, "both in the information that it uses and in the resources that it requires for processing that information" (Carruthers, 2006, pp. 57–9). Low-level perceptual mechanisms, for example, are encapsulated because they do not take into account most of the agent's relevant beliefs or desires during task performance.

Spelled out in more detail, information O is encapsulated from module M when three conditions are met: [43]

1) O is relevant to M's success in a task drawn from M's domain.
2) M fails to use O in performing the task, despite the fact that O is available and some other systems reliably use it to perform tasks.

---

[40] Some MM theorists are committed to empirical adaptationism, and so often mistakenly assume that adaptations are optimal designs in some sense (Godfrey-Smith, 2001). Here, I make a more reasonable claim of satisficing performance, which also connects better with my characterization of Darwinian modules as implementing Darwinian heuristics.

[41] Hence, central systems, as I discussed earlier, are domain-general in the sense that they are responsible for tasks in, and receive inputs from, a wide range of domains.

[42] It is controversial whether information encapsulation defined this way is too weak to be interesting (Samuels, 2006). However, I will grant it to MM for the sake of giving them as many theoretical resources as possible.

[43] This characterization of core conception is adapted from Fodor's recent definition of informational encapsulation (Fodor, 2000, p. 62).

3) The inaccessibility is due to "architecturally imposed" features of M.[44]

The notion of informational encapsulation adopted here is what Carruthers calls "wide-scope encapsulation". According to Carruthers, there are two types of informational encapsulation: narrow-scope and wide-scope (Carruthers, 2006, p. 58).[45] A narrow-scope encapsulated system is such that, for most information held in the mind, the system's operation cannot be influenced by that information. That is to say, there is a fixed and small subset of information this system can utilize. In contrast, wide-scope encapsulation captures a more flexible notion of a module. A wide-scope encapsulated system is such that it cannot be influenced by most of the information held in the mind during the process of a task. There is no fixed division, however, between information it can and cannot utilize across different tasks.[46]

There are several ways informational encapsulation can be realized. According to the most influential notion characterized by Fodor (1983), a computational mechanism can (and, according to Fodor, has to) achieve informational encapsulation through possession of the following features: (1) a proprietary database, (2) extremely limited access to information outside of the database, and (3) insensitivity to higher-level feedback information. As the information it can utilize is largely confined to what is in its own database, the computational mechanism cannot access the otherwise vast amount of available and relevant information. However, Fodor's proposal only represents one extreme and demanding way of realizing informational encapsulation, and few empirical psychologists think of information encapsulation this way. Interestingly, even Fodor has abandoned this proposal for the following two characterizations, which are much more liberal (Fodor, 2000, p. 62):

First, informational encapsulation can be achieved by controlling factors external to the mechanism-for example, restricting the information flow (Fodor, 2000, pp. 62–3). The encapsulated mechanism can be connected to the other mechanisms in the cognitive system in a way that its access to the encapsulated information is restricted. For example: hypothetically, a face recognition module is constructed so that it can make use of auditory representations if they are made accessible to it; however, because the module is only connected to other vision-related mechanisms, information from auditory systems has no way of reaching it. The face recognition module is thus informationally encapsulated from auditory information through the restriction of information flow. Note that this way of restricting information flow is less demanding than the earlier notion of

---

[44]Here, I follow Samuels' understanding of architecturally imposed feature: a feature is architecturally imposed minimally implies that (1) it is a relatively enduring property of the system, (2) it is not a product of performance factors, e.g., motivation, preference, attention, resource, etc., and (3) it does not change as a result of alternation of representational states of the organism (beliefs, goals, etc.) (Samuels, 2006, p. 39).

[45] Narrow-scope and wide-scope encapsulated mechanisms are what Samuels calls diachronically and synchronically encapsulated mechanisms respectively (Samuels, 2005, p. 112).

[46] Samuels argues that wide-scope encapsulation, with its lack of fixed set of encapsulated information, is not an interesting notion of informational encapsulation for two reasons. First, it is different from what most theorists mean by "encapsulation." Second, any mechanism that does not run exhaustive searches would qualify as wide-scope encapsulation, and because almost no one thinks exhaustive search is characteristic of human cognition, wide-scope encapsulation cannot be a distinctive and interesting notion (Samuels, 2006, p. 45). I must confess that I do not find any of the reasons convincing. However, it is not a place to offer arguments (Carruthers (2007) offers an argument for the legitimacy of the notion of wide-scope encapsulation). Here, I will simply assume it is a legitimate notion for the purpose of arguing against evolutionary psychology. Allowing for the resource of wide-scope encapsulation will only make it harder, not easier, for me to argue against MM.

information encapsulation. Because it does not require a fixed proprietary database, it potentially allows an encapsulated module to utilize a larger and more flexible range of information.

Second, a mechanism can achieve informational encapsulation through internal design. It can implement algorithms that are formulated in such a way that they do not apply to the information the mechanism is encapsulated from, even if it is made accessible (Fodor, 2000, pp. 62–3). For instance: hypothetically, a face recognition module can achieve informational encapsulation by implementing an algorithm that does not apply itself to any auditory representations. Or, hypothetically, a reasoning module can be encapsulated by implementing a salience-sensitive heuristic that only takes the three most salient pieces of information into considerations. Finally, it is important to note (for the discussion in Section 2.3 of Chapter 3) that informational encapsulation achieved this way does not rely on any form of restriction on the flow of information, as the previous two do. All relevant information can potentially be made accessible and flow through a module without the majority of the relevant information being utilized by the module when it performs a task. As such, encapsulation does not imply restriction on information flow.

The first way of realizing encapsulation discussed above only produces narrow-scope encapsulation, due to the limited access to information in the proprietary database. However, the second and third ways can produce either narrow-scope or wide-scope encapsulations, depending on the architectural details. In short, Darwinian modules are informationally-encapsulated either in a wide-scope or in a narrow-scope sense. They do not utilize all the task-relevant and available information in information processing.

Innateness

Many Darwinian modules are innate in the sense that, roughly, they are not the result of learning (Cowie, 1998; Samuels, 2002, 2009).[47] This definition of innateness is explicitly endorsed by some MM theorists (Barrett, 2005b; D. Sperber, 2007). It is also compatible with MM theorists' recent advocacy of an interactionist account that does not "see nature and nurture as existing in an explanatory zero-sum relationship… all properties of the organism equally develop through 100% gene-environment interaction" (Tooby & Cosmides, 2016, p. 34). Finally, this definition of innateness is compatible with the commitment nativists ought to uphold that biological evolution is more powerful than individual learning and cultural evolution in producing adaptive traits (Boyd & Richerson, 2007; Jeffares & Sterelny, 2012).

To sum up, according to MM theorists, the human mind is a computational system consisting entirely in massive numbers of domain-specific, informationally encapsulated, computational mechanisms that are innate or the product of learning.

5. The Massive Modularity Hypothesis: Epistemic Commitment

Characterizations of MM often focus on its architectural commitment to Darwinian modules. In this section, I pay special attention to MM's epistemic commitment—the heuristic approach to human problem-solving. My aim in doing so is to clarify and highlight the epistemic property of modules.

---

[47] According to the psychological primitiveness accounts of innateness, innateness in the context of cognitive science refers to properties of a cognitive system that (1) emerge in the course of normal development and (2) admit of no psychological explanation (e.g., learning-based explanation) (Cowie, 1998; Samuels, 2002, 2009). Innate properties are "cognitive primitives"—they are appealed to in the explanation of other mental phenomena, but they themselves do not admit of any psychological explanation (as opposed to a biological one).

This epistemic property of heuristics has not received sufficient attention in the debate between MM theorists and their opponents despite how crucial it is to understanding MM's limitations in explaining behavioral flexibility. I conclude this section by making explicit the relations between MM's epistemic and architectural commitments.

5.1. The Heuristic Approach to Reasoning

MM's epistemic commitment is the heuristic approach to reasoning (Carruthers, 2007). This approach develops out of bounded rationality, a broader research tradition. This tradition takes the empirical limitations of human cognition, such as cost of information and computational resources, into consideration and offers a more plausible account of human rationality (Simon, 1955, 1956). The heuristic approach claims that human reasoning competence consists of a set of rules and principles that should be properly called heuristics.[48]

It is hard to define "heuristic," a term that is extremely popular in divergent disciplines. It is used in different ways, and its meaning has evolved over the years (Gigerenzer & Todd, 1999). For instance, the concept of heuristic plays a central role in a few disparate but associated research programs, such as Herbert Simon's information processing theory and models of artificial intelligence (Simon & Newell, 1958), Kahneman and Tversky's heuristics and biases research program (Kahneman, Slovic, & Tversky, 1982), and Gigerenzer's fast and frugal heuristics research program (Gigerenzer, 2004). Nevertheless, we can settle for a minimal characterization, a set of selective central features of heuristic that is good enough for our purpose:

> Heuristics are principles or algorithms for information processing that provide (1) satisficing outcomes, (2) within the constraint of tractable computation.[49]

Let me illustrate these two features: First, satisficing outcomes are in some sense "good enough" relative to a standard determined by the relevant context. Consider the "Do-what-the-majority-do heuristic": If you see the majority of your peers display a behavior, engage in the same behavior (Gigerenzer, 2006, p. 126). Studies have reported these imitation behaviors in both humans and non-human animals. For instance, female guppies base their mate choice on the preferences of other female guppies (Dugatkin, 1992). Such simple social heuristics seem to result in adaptive behaviors in many cases (Laland, 2002). Adaptive behaviors are satisficing solutions but need not be optimal solutions.

---

[48] The heuristic approach should be distinguished from the approach of "optimization under constraint". The latter research program retains the pursuit for optimality while taking into account the costs of time and resources human cognitive system pays. For example, in a version of rational theory of memory (J. R. Anderson, 1990, p. 46), search for an item in memory continues until the expected benefits of further search no longer exceed the expected costs. Optimality is retained in the account: "the stopping point is the optimal cost-benefit trade-off" determined by complex cost-benefit calculation (Gigerenzer, 2006, p. 117), and the solution is "optimal under constraint". However, such approach faces the same problems of the Standard Account—the computation required for the solution involves intractably complex computation (Gigerenzer, 2006, p. 117) and it is also psychologically unrealistic (Richardson, 1999, p. 569).

[49] This definition of "heuristic" is based on Sheldon J. Chow's (2015) comprehensive treatment of heuristics. However, my definition is weaker than Chow's—heuristics are not required to be processes involving representations, they can instead be purely causal processes. I remain neutral as to the best way to define 'heuristic' for the purpose of scientific studies, which is Chow's main concern. This weaker definition captures how MM theorists and their allies use this term (Gigerenzer & Gaissmaier, 2011; Todd & Gigerenzer, 2012).

Second, heuristic processes are computationally tractable. They are "fast and frugal" as they operate quickly and require little information and computational resources. To achieve tractability, heuristic computational processes deviate from ideal ones in roughly two aspects. First, they take less than all relevant information to bear on the computation. For example, they look into only a small subset of available options. They do not run exhaustive searches for all relevant information, and only take a selective subset into consideration. Second, they use principles or rules that are computationally less complex than ideal theories. For example, they may dispense with the Bayesian calculation of conditional probabilities.

In fact, computational tractability is made possible because heuristics trade optimal solutions for satisficing ones, and applicability in universal contexts for selective ones. Because heuristic processes aim to produce only satisficing solutions, they do not need to consider all options but can simply stop computing once they find an option that satisfies the relevant standard. Also, because heuristic processes do not need to make an exhaustive utility ranking of all options to find out the optimal one, they can dispense with the complicated exhaustive computation of probabilities and utilities that is necessary for such tasks.

More importantly, because heuristic processes are specialized in selective contexts, they can build in "knowledge" about their specialized contexts.[50] In turn, such knowledge can save them from a lot of complicated computation. For example: if an information structure in a particular context is such that a piece of information A is always correlated with a more easily computed information B, the specialized heuristic process that "knows" about this can use the structure to its advantage. For example, it can replace a complex computation for information A with a more tractable computation for information B. In addition, if a heuristic process "knows" that in its specialized contexts certain types of information are rarely relevant, it could then ignore them in its computation with little risk of substantial consequences. As a result, it need not search for the information nor compute to determine its relevancy. Consider the "recognition heuristic": if you recognize only one of the two options, infer that the one you recognize has a higher value of X in question (Gigerenzer, 2006, p. 124). For example, if you only recognize the name of one city but not another, it is very likely that the city you recognize has a larger population. This heuristic works in an environment where the recognition validity is larger than chance (0.5), but does not provide correct answers reliably enough otherwise.[51] The "recognition heuristic" achieves tractable computation by assuming a correlation between a city's size and the recognizability of its name. In deciding which city has a larger population, it does not take into account all relevant information but only considers which city evokes recognition.

As a result, for any given heuristic, the quality of its solution is relative to its context of performance. That is, the performance of a heuristic is context-dependent: while it reliably provides satisfying solutions in some contexts, it does not do so in others. "[A] heuristic is just a procedure that doesn't always work…," as Fodor (2008) puts it. Or, more sophisticated in Gigerenzer's words (2006, p. 121), "…a heuristic is not good or bad, rational or irrational per se, but only relative to an environment." Note that the degree of specialization and the degree of computational tractability are positively correlated. The more specialized a particular heuristic is, the more it can assume about its specialized contexts, and the more relevant information and

---

[50] The knowledge can be either represented explicitly in the heuristic rules or principles, or built-in implicitly in the organization or derivation of those rules.

[51] The recognition validity is defined as the proportion of cases where a recognized option has a higher value of X than the unrecognized options (Gigerenzer, 2006, p. 124).

computation it can dispense with. However, such trade-offs come with a price. Because the built-in "knowledge" is only true of its specialized contexts, when the heuristic process is applied to contexts other than those, its performance soon plummets.[52]

In conclusion, MM's epistemic commitment is the heuristic approach to reasoning. By nature, heuristics provide satisficing solutions with the benefit of tractable computation. However, heuristics' performance is context-dependent. This is an important point we will come back to in the next chapter.

5.2. The Relationship Between the Epistemic and Architectural Commitments

We are now in a position to see the relationship between MM's epistemic and architectural commitments: Darwinian modules (MM's architectural commitment) mostly[53] implement heuristics (MM's epistemic commitment). Relatedly, the heuristics implemented by Darwinian modules (a.k.a. *Darwinian heuristics*) are domain-specific—they are designed to solve a limited range of tasks (Carruthers, 2007; Fodor, 2000).[54] Let me elaborate more on this relationship.

To begin with, one core feature of heuristics is that they do not take into account all relevant information available in the cognitive system; informational encapsulation, however achieved, implements this feature. For example, hypothetically, a recognition heuristic can be built into a computational mechanism by (1) limiting the mechanism's information access to a proprietary database containing nothing but recognition information, (2) constraining the information distribution so that little beyond recognition information can be accessible to the mechanism, or (3) implementing an algorithm that ignores almost everything but recognition information.

Moreover, both heuristics and modules are limited in their applicable contexts in some way. As we have seen, the performance of heuristics is context-dependent—they only work satisficingly in selective contexts; Darwinian modules are domain-specific—they are designed to perform selective tasks. It is worth noting that context-dependence and domain-specificity are two related but distinct features that MM theorists have not clearly distinguished. As we will see, this is one of the main reasons they fail to recognize their inability to explain behavioral flexibility. We will come back to context-dependence and domain-specificity in the next chapter when I discuss the nativist information control problem.

---

[52] This tight match between a heuristic and its task environment is a founding principle behind ecological rationality, according to Gigerenzer:

> "Human rational behavior by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor" (Simon, 1990). Just as looking at one blade of a scissors would not provide insight on how it cuts, one cannot understand fully how the human mind works by ignoring the task environment (Gigerenzer, 2004, p. 67).

[53] I say "mostly" because Darwinian modules may also implement simple non-heuristic algorithms, such as the logical rule of modus ponens. The main contrast here is between heuristic and optimal non-demonstrative inference algorithms. Here, I will ignore a complication that Darwinian heuristics can be implemented by domain-general mechanisms, such as in the Library Model of Cognition (Samuels, 1998a).

[54] In order to make sense of domain-specific heuristics, we need to distinguish the general "type" and the specific "token" of a heuristic. A given type of heuristic may be selected over and over again to be implemented as specific tokens of information processing algorithms for specific tasks in different domain-specific modules (Carruthers, 2007).

In sum, the massive modularity hypothesis—the view that the human mind is composed exclusively of a massive set of Darwinian modules—represents one special way of developing the heuristic approach into an account of cognitive architecture.[55] MM can be seen as a reaction to the Standard Account's aforementioned shortcomings: we cannot possess the ideal non-demonstrative reasoning competence that would otherwise explain our behavioral flexibility. Thus, MM will have to provide an alternative account of our flexible problem-solving capacity in a wide range of novel contexts, one that is not based on the heuristic approach to human reasoning.

6. Conclusion

In this chapter, I have analyzed the explanandum of MM architecture—the intuitive conception of behavioral flexibility that humans have the competences to perform tasks satisficingly in a wide range of significantly different novel contexts. I also introduced the explanans—the massive modularity hypothesis, according to which the human mind is composed exclusively of a massive set of Darwinian modules. In particular, I defined the key concepts involved in MM in a charitable way in order to provide MM as many theoretical resources as possible to deal with some existing criticisms against it, which will be discussed in the next chapter. I also contrasted MM with the Standard Account of cognitive architecture in order to highlight MM's theoretical motivation and epistemic commitment, i.e., the heuristic approach to reasoning. However, as I will show in the next chapter, this epistemic commitment is not adequately appreciated by many MM theorists and their opponents. Having clarified the key concepts, I am now better placed to reveal an important gap in MM's explanation of human flexibility, and to show why MM has no resources to fill it.

[55] Other cognitive architectures that implement the heuristic approach include (but are not limited to): a cognitive architecture that combines an epistemic commitment of heuristic approach and an architecture of central system (Miller et al., 1960; Newell & Simon, 1972; Samuels, 2005, 2010, 2010).

3

Massive Modularity and the Nativist Information Control Problem II: The Explanatory Gap

1. Introduction

In the last chapter, I defined and clarified the explanandum—the intuitive conception of behavioral flexibility, and explanans—the massive modularity hypothesis (MM). According to the intuitive conception of behavioral flexibility, humans have the competences to perform tasks satisficingly in a context-appropriate fashion, in a wide range of significantly different novel contexts. For MM to explain this flexibility, it will need to show how a massive set of Darwinian modules, each of which is designed to solve a limited range of adaptive problems, can work together to generate context-appropriate solutions in a wide range of significantly novel contexts.

In this chapter, I will show that there is a significant gap in MM's explanation of behavioral flexibility, and that recent MM models, despite their incorporation of learning, development, and self-organization, still have no theoretical resources to fill this gap. This is because in order to explain human flexibility, it is necessary that MM addresses the information control problem, by showing how information can be routed to the relevant modules in order to support a specific task. However, MM faces what I will refer to as the *nativist information control problem*: Darwinian modules cannot (learn to) perform control reliably in a wide range of novel contexts where human reason thrives. In short, nativist information control problem is why MM still cannot explain human intelligence.

In Section 2, I will briefly review the most recent MM models and explain how they overcome Fodor's challenges. Having established that these objections are answerable, I go on to reveal a serious explanatory gap for the MM model, which indicates that MM is unable to solve the control

problem nevertheless. To prove that MM lacks the theoretical resources for flexible information control, I will first consider an extreme nativist version of MM (Section 3), before turning to a moderate version (Section 4). My arguments will make it clear that no amount of self-organization, development, and learning can help MM overcome the control problem, as long as it remains committed to nativism. One insight will become increasingly clear: one cannot get a flexible whole if one limits the flexibility of all interacting parts. Section 5 will respond to potential objections to my arguments. In Section 6, I close the chapter with a brief discussion of some broader implications for control mechanisms and cognitive architecture.

2. The Explanatory Gap

In this section, I will expose the explanatory gap. First, I will introduce a recent MM model incorporating new theoretical resources of self-organization and learning. Then, I will motivate the control problem and show that it poses an explanatory gap this model needs to fill. Finally, I review Fodor's two challenges that MM fails to fill this gap and discuss replies from MM theorists that his criticisms miss the mark. The next two sections will advance two new arguments that MM indeed has no theoretical resources to fill this gap, but not for reasons Fodor thinks.

2.1. The Confederate Account of Massive Modularity

Human flexibility seems to pose a special challenge to MM, because special-purpose cognitive mechanisms that evolved for adaptive problems do not inspire confidence in their capacity to solve novel problems. Historically, philosophers have made this point multiple times:

> [E]ven though … machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they were acting not through understanding but only from the disposition of their organs. For whereas reason is a universal instrument which can be used in all kinds of situation, these organs need some particular disposition for each particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act. (Descartes, 1984)

And more recently:

> It is …. not plausible that the mind could be made only of modules; one does sometimes manage to balance one's checkbook, and there can't be an innate, specialized intelligence for doing that. (Fodor, 1998, p. 155)

As a result, new MM models are committed to what Samuels (2012) calls a "confederate account," according to which flexibility is the product of "a network of subsystems that feed each other in criss-crossing but intelligible ways" (Pinker, 2005: 17).

However, for this account to be plausible, MM theorists need to show that their models can manage the control of information reliably. Specifically, they need an account of how the right information will be routed to the relevant modules at the right time. Without a solution, the confederate account is not "an explanation of our cognitive-behavioral flexibility so much as a statement of the problem given a commitment to MM" (Samuels, 2012: 80). Here is where some recent models come in, which aim to show that they have resources to achieve self-organized control of information (Barrett, 2005a; Carruthers, 2006; Sperber, 1994; Sperber & Hirschfeld, 2006). For the sake of brevity, I will provide only an overview of Carruthers's (2006) original and well-developed model, but I will point out incorporated contributions from other authors:

1. **Massive modularity:** The human mind is constituted by perceptual, motor, and central (i.e., belief- and desire-generating, and practical reasoning) modules (*Figure 3.1*).

2. **Interactive control:** Perceptual information is broadcast to belief- and desire-generating modules through a bulletin board structure (*Figure 3.2*). Each module has a recognition front-end that identifies representations that fit its triggering conditions. If a representation satisfies its triggering conditions, it will be activated to process the representation. If there are multiple representations meeting its condition, they compete to become its inputs. Other modules and representations may also modulate the module's activities at the same time. In this way, interactive control works like enzymes interacting and influencing the production of proteins in the cells (Barrett, 2005a). We might say that the bulletin board structure, the recognition front-ends, and the interactions between modules all work together as a soft-assembled control mechanism (Clark, 1998). Such soft-assembled interactive control mechanisms can interface with other types of modules as well.

3. **Learning:** In response to the demands of, for example, a physical problem, some motor plan is activated. If the problem can be solved by performing the action, the motor plan will be executed. If otherwise, learning processes will generate variations in various modules (including control modules) to produce a different motor plan.

4. Step 3 can be repeated until a satisficing solution emerges or until the agent gives up.

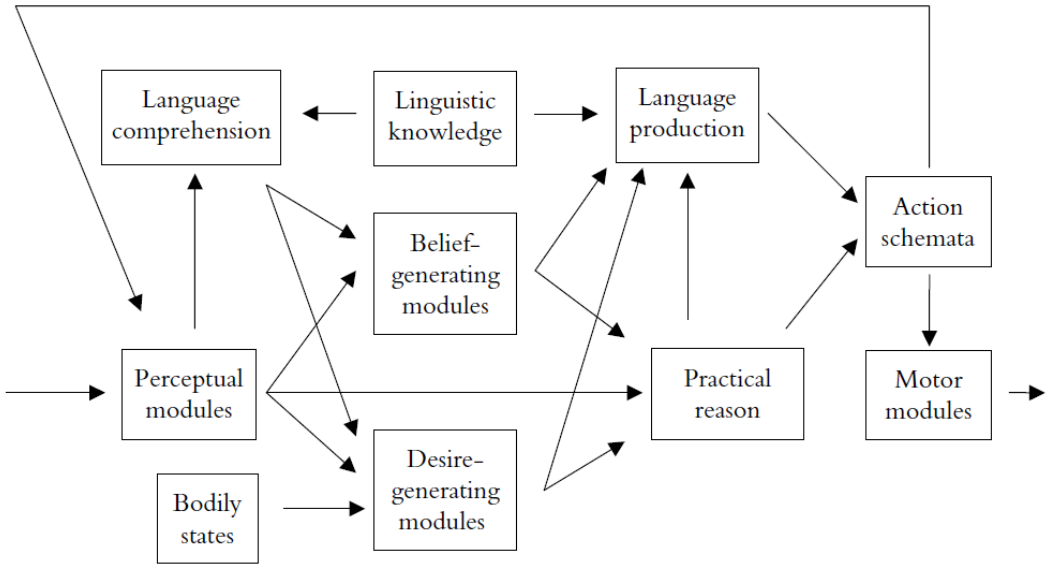

Figure 3.1 Carruthers's massively modular architecture of the mind. Each box represents multiple modules. Excerpted from (Carruthers, 2006).

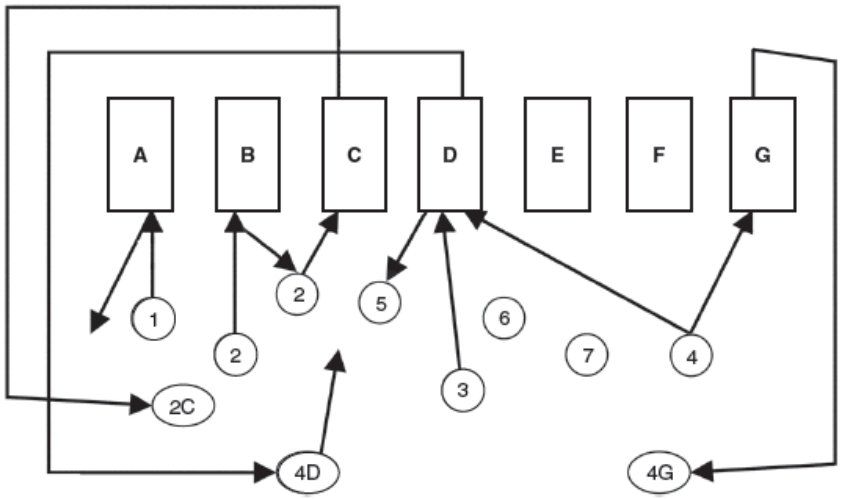

Figure 3.2 The bulletin board architecture. Modules (labeled A through G) constantly monitor representations (labeled 1 through 7) broadcasted to the bulletin board and process those that fit their triggering conditions. Excerpted from (Barrett, 2005a)

2.2. The Information Control Problem

This account indeed looks plausible. Soft-assembled control mechanisms potentially allow more flexible control than control modules do. The cycles of mental operations with a learning capacity promise to assemble solutions to novel problems from a preexisting "adaptive toolbox" of well-designed modules (Gerd Gigerenzer & Selten, 2002). However, there is a significant explanatory gap. To identify it, let me lay out the argument implicit in the model:

(1) To perform satisficingly in a context, Darwinian modules will be assembled properly so that the relevant modules are engaged at the right time during information processing.

(2) To assemble the relevant modules properly, control needs to be performed at (almost) each step of the information processing.

(3) Darwinian modules (or the soft-assembled control mechanisms)[56] can (learn to) perform control satisficingly at each step of information processing for a wide range of novel contexts.

(C) MM models can perform satisficingly in a wide range of novel contexts, i.e., MM can explain flexibility.[57]

Premise (3) presents us with the explanatory gap: Despite implicating complex interaction and learning, MM theorists have not demonstrated its truth. Let me motivate this explanatory gap further. Control modules need to determine which downstream module is relevant, and this is essentially a decision concerning what context one is currently in. However, context-determination cannot be taken for granted, because it usually involves non-demonstrative inference, which is

[56] Here, I assume the soft-assembled information control mechanisms to be functionally equivalent to Darwinian information control modules. Section 5.1 will address an objection to this assumption.

[57] Here, I will grant the truth of a hidden premise: all the building blocks (non-control modules) needed for assembly are available either innately or through learning.

isotropic: it may depend on any arbitrary piece of information, given appropriate background beliefs.[58]

For instance, MM suggests there is a cheater detection module (Cosmides & Tooby, 2013). In order to route relevant information to this module, a control module needs to make the decision that one is currently in a social exchange situation. However, social exchange is not marked by any invariant sensory cue. As Fodor nicely puts it: "Even if all the social exchanges used to be colored orange, so that Way Back Then the [cheater detection module] could identify its inputs by their color, it surely can't do so these days" (2000, p. 76). However, despite potential initial errors, we have no problem interacting socially via the most impersonal computer interface; we can also interact non-socially when engaging an (asocial) robot that puts on a human face.

A different example involves reading, a novel skill for which there cannot be evolved modules. According to Barrett and Kurzban, MM architecture may subserve reading by recruiting the existing module for "object recognition… then linking them to systems for naming, semantics, and so on" (2006, p. 638).[59] However, for reading to work, a control module will need to acquire the competence to make the correct decision in the novel context of reading and route the outputs of object-recognition module to the naming module.

Two important criteria for control modules thus result: (1) the ability to take information potentially from anywhere into consideration, and (2) the relevant competence to make the correct decision about contexts. So, the crucial question is: can MM's control modules meet the above two criteria? The answer is "No," according to Fodor: He raises two "input problems" against MM's ability to handle control adequately (Fodor, 2000). I will now briefly review his arguments, as well as their solutions. The goal is to show that the nativist information control problem that I reveal is conceptually distinct from Fodor's input problems, and that the nativist information control problem is a problem MM lacks resources to solve.

2.3. *The A Priori* Input Problem

The first input problem is introduced by Fodor as an "*a priori* argument" against MM. To put the argument more straightforwardly:

(1) In order to route information to a module M, a less modular upstream control mechanism is required.

Because the control mechanism needs to route, from a more general pool of information, appropriate inputs to module M, it must take in a wider range of information as inputs and, according to Fodor, be less domain-specific than its downstream module M (Fodor, 2000). This is because Fodor defines domain-specificity as limitation to a module's task-domain *as well as* input-domain—the types of representations a module accepts as input. Furthermore, this control mechanism will itself require an even less domain-specific upstream control mechanism. Hence,

---

[58] As a result, the only way to take all relevant information into account is to actually assess (almost) all of the available information in one's background beliefs. This is why optimal non-demonstrative inferences require intractable computation. This also relates to the infamous frame problem (Fodor, 1983; Shanahan, 2016).

[59] I will grant that the object-recognition module, although not designed for reading, can take words as inputs, i.e., words are the "actual input" but not the "proper input" for the module (Dan Sperber, 1994); also, I will grant that the object-recognition module have the relevant competence to process words.

(2) It will come to a point where a non-modular cognitive mechanism is required to do the routing. As a result,

(C) An architecture composed exclusively of modules can handle control only by positing increasingly non-modular structures, which in the end isn't a massively modular architecture

Many arguments have been put forward against the *a priori* input problem (Barrett, 2005a; Carruthers, 2006; Clarke, 2004; Collins, 2005; Deise, 2008; Shanahan & Baars, 2005; Dan Sperber, 1994; Weiskopf, 2002; Wilson, 2004). Because they all provide similar solutions, here I will focus on their commonality. The trouble with the *a priori* input problem is that premise (1) is false: Because MM defines domain-specificity only in terms of limitation to task-domain, a control module is no less modular than its downstream modules even if it has access to a wider range of information. In short, the *a priori* input problem fails because it attacks MM based on a non-essential architectural feature (i.e., limitation to input-domain).

2.4. The Really Real Input Problem

Fodor raised a second input problem that is, according to him, not just philosophical but "really real" (Fodor, 2000, p. 77). The argument is as follows:

(1) To solve a cognitive task with a heuristic, an additional "routing" decision is required to figure out the identity of the present context so that the appropriate heuristic can be selected to perform the task.

(2) Heuristics cannot reliably perform non-demonstrative reasoning.

Because heuristics trade off optimality with tractability, they cannot be optimally reliable in all contexts.

(C) Therefore, MM model cannot perform control reliably.

The problem of this argument is its premise (2): MM theorists contest its truth by taking issue with the conception of reliability involved (Jackendoff, 2002; Pinker, 2005). Premise (2) is true only if the conception of reliability involved is one stronger than the satisficing type of reliability associated with heuristics. However, there is no reason to adopt a stronger and unrealistic conception of reliability—not if our aim is to explain human flexibility at least. In short, the "really real" input problem fails because it employs an unrealistic epistemic standard (optimality across contexts) in criticizing MM.

I have argued that recent MM models have a significant explanatory gap: they leave information control unexplained. Whereas Fodor sets his sights on MM's architectural and epistemic commitments in his criticisms of the hypothesis, I will argue in the next two sections that the core of MM's control problem lies in its nativist commitment. The nativist information control problem, as I shall call it, is what MM particularly lacks resources to solve.

3. The Nativist Information Control Problem for an Extreme Version of Massive Modularity

In this section, I will argue against an extreme version of MM, according to which all Darwinian modules are innate. It is important to stress that few endorse this view. Nevertheless, it is instructive to consider it because it will clearly highlight the consequences of the nativist commitment: due to the extreme nativist commitment, Darwinian modules can't perform control

satisficingly at each step of information processing for a wide range of novel contexts—i.e., the nativist information control problem is the reason MM cannot fill the explanatory gap. As I will show, Darwinian modules are likely to perform control non-satisficingly at least at one step of information processing in a novel context. MM theorists fail to see this consequence because they have not clearly distinguished the domain-specificity and context-dependence of Darwinian modules.

Two preliminaries before the argument. First, if a context is novel, the control modules involved to provide a solution will encounter novelty at one of the steps in information processing, if not more. For example, for the cheater detection task discussed above, there may be nothing novel in the perceptual and motor processes—presumably, our perceptual and motor interactions with the world remain largely the same. However, the *cognitive* processes for determining the context of social exchange would have changed tremendously.

Second, a heuristic's performance is context-dependent. Context-dependence is a means to achieve its tractable and satisficing performance. Because a heuristic builds in competence for specific contexts, it "knows" whether a piece of information is likely to be relevant, or it can be approximated or ignored without significant loss in performance in those contexts. As a result, it can perform well in those contexts and becomes unreliable in others. That is why, "… a heuristic is not good or bad, rational or irrational per se, but only relative to an environment" (Gigerenzer, 2006). Nevertheless, we must observe that by committing to heuristic information processing, MM introduced some form of inflexibility to all the modules in its model—they have competences for satisficing performance only in specific contexts.

For example, people and other animals sometimes adopt the recognition heuristic for food selection tasks: we prefer food that we can recognize (Heyes & Galef, 1996). This heuristic is satisficing (not all recognizable food is optimal for us) and tractable (this strategy requires minimal cognitive resources), but it only works well in specific contexts (where the qualities of different types food remain relatively stable so that information about their qualities can be socially transmitted reliably).

Importantly, while context-dependence concerns the fit between a module's competence and the relevant contexts, domain-specificity concerns the proper function of a module. We can see this distinction better by noting that while domain-specificity is determined by causal-historical facts (whether it *was* designed by evolution for certain tasks), context-dependence is not (the performance simply *is* or *isn't* satisficing in a particular context). Also, while domain-specificity concerns only the task-domain, context-dependence concerns the entire contexts (the information structures of tasks in environments). For example, a cheater-detection module is domain-specific because it was designed to only perform the cheater-detection task. It is also context-dependent because it embodies competence that only works satisficingly for the cheater-detection task in specific environments, such as a default assumption that an agent's behavior is caused by internal mental states.

Because of the significant difference between the two features, the inference from a module's task-domain to its actual performance is complicated. The inability to see this underlies the failure of MM to appreciate the nativist information control problem. In particular, MM theorists make the illegitimate inference from the control module's proper function to route information on the one hand, to its satisficing control performance in a wide range of significantly different contexts on the other. Hence, there seems to be no explanatory gap for them. However, recall the recognition heuristic for food selection above. Although it performs satisficingly in stable environments, its performance plummets in unstable ones.

Now we are ready for the argument. In the following, I will show step-by-step that although we can make an inference (i) from a Darwinian heuristic's proper function of task-domain T to its likely satisficing performance of T *in an adaptive context*, we *cannot* make an inference (ii) from its proper function of T to its likely satisficing performance of T *in a significantly novel context*. In fact, the inference we should make is (iii) from its proper function of T to its likely *non-satisficing* performance of T in a significantly novel context. As a result, we should conclude that Darwinian control modules are likely to perform control non-satisficingly in a wide range of significantly novel contexts. Let us examine this argument more closely.

Uncontroversially, we can make the inference (i) from a Darwinian heuristic's proper function of task-domain T to its likely satisficing performance of T *in an adaptive context*. This is because evolution by natural selection is, given a strong selection force, a reliable process for producing adaptive traits for solving adaptive problems in the EEA. The adaptive traits are built incrementally through generations of selective retention of variation in heritable traits. Note that if a randomly-generated variation is retained reliably due to its function, it is because the variation has higher fitness through solving adaptive problems *in its selective environment*, the EEA. That is, evolution can only generate adaptive traits, the utility of which are highly relative to the context under which they are selected. In the case of a control module, the adaptive trait includes the built-in competence for control, which enables the module to control information routing satisficingly with respect to contexts present in the EEA.

However, we *cannot* make the inference (ii) from a Darwinian heuristic's proper function of T to its likely satisficing performance of T *in a significantly novel context*. This is because the competence required for satisficing performance in an adaptive context is very different from that in a significantly novel context. MM theorists understand the importance of fit between competence and information structures for generating satisficing solutions for different types of tasks:

> …what counts as a solution differs radically and incommensurably for different adaptive problems. Consider, for example, food choice versus mate choice. The computational structure of programs that are well engineered for choosing nutritious foods will fail to produce adaptive behavior unless they generate different preferences and tradeoffs than programs designed for choosing fertile sexual partners. (Cosmides & Tooby, 2013, p. 203)

However, what they forget is that the same type of task, when in different environments, can have very different information structures. Specifically, because task-relevant information varies across contexts, the built-in competence, outside of its adaptive contexts, can mislead the heuristic into considering irrelevant information, ignoring relevant information, and approximating crucial information with an inappropriate algorithm. For instance, the built-in competence in our visual system assumes light comes from overhead and can result in illusory perception in environments that violate this assumption (Scholl, 2005). As a result, the competence useful for control in the EEA is very likely to be detrimental for routing information in a significantly novel environment.

Consider the control module involved in routing information to the cheater detection module. MM theorists take for granted that the control module can route information reliably both in the modern world and in the EEA. However, because many modern social exchange situations (e.g., interacting via an extremely impersonal computer interface) are significantly different from the ones in Pleistocene, the built-in competence designed for the latter is likely to be unhelpful for the former. That is, we cannot infer from the control module's proper function to its likely satisficing performance to route information in those significantly different social exchange situations.

Of course, it is possible that the built-in competence, although not designed for a novel context, turns out to be helpful. After all, evolution does not design its work to prevent other possible

applications. However, this possibility remains thin because a good fit between evolved competence and the specific novel information structure confronting the organism can happen only by chance, but cannot result from some reliable processes. As a result, we cannot make the inference from a module's proper function of control information to its *likely* satisficing control performance in a significantly novel context.

Instead, we should make the inference (iii) from a Darwinian heuristic's proper function of T to its likely *non-satisficing* performance of T in a significantly novel context. This is because evolution is *not* a reliable process for building competence useful for a future novel context: "Evolution cannot prepare the organism for what is to be, only what was." (Barrett & Kurzban, 2006, p. 635). A variation of built-in competence that is useful *only* for a future novel context could not be reliably retained in the EEA because it would not increase one's fitness at the time of selection. Indeed, to be reliable in producing traits useful for novel contexts would require of evolution the kind of foresight that it cannot possibly have: The ability to predict, thousands of years ago, the novel environments the human species will encounter and the fitness benefits a variation will confer, and to retain that variation. In short, we can only infer from a module's proper function of control to its likely *non-satisficing* control performance in a significantly novel context, without violating our fundamental conceptions of the evolutionary process.

In the case of reading, for example, MM theorists maintain that information from the object-recognition module (recruited to perform a novel task) can be routed to the naming module to subserve the reading process. However, as reading is a novel skill that does not exist in the EEA, relevant control competence cannot be produced by evolution reliably. As a result, we can only infer that the relevant control module will likely perform routing non-satisficingly in the context of reading.

Now, if we should infer from a Darwinian control module's proper function to its likely (although not necessarily) non-satisficing performance *in a significantly novel context*, we certainly should infer that it will not be satisficing *in a wide range of significantly novel contexts*. Additionally, the likelihood of failing control gets compounded as Darwinian control modules need to manage more than one step of control involved in a complex task.

In short, I've shown that the extreme version of MM suffers from the nativist information control problem and as a result, cannot explain human flexibility. Stepping back from the details of the arguments, one can perhaps see more clearly why the explanatory gap cannot be filled: To build a very flexible self-organized system, we simply cannot have all its components limited in their flexibility. MM attempts to do exactly this with its commitments to both heuristics (inflexible due to the context-dependent performance) and to nativism (inflexible due to the limitation to learning).

4. The Nativist Information Control Problem for a Moderate Version of Massive Modularity

So far, I have focused on the extreme version of MM that is incompatible with learning. However, other versions of MM, though a nativist program, are compatible with it (Barrett & Kurzban, 2006; Carruthers, 2007; D. Sperber & Hirschfeld, 2006; Tooby & Cosmides, 2016). In fact, many MM theorists embrace a moderate version and can happily acknowledge that, "most innate cognitive modules are domain-specific learning mechanisms that generate the working modules of acquired cognitive competence" (D. Sperber, 2007, p. 57). So, despite the fact that the extreme version of MM may have ignored a significant difference between domain-specificity and context-dependence, and therefore invited wrong inferences, learning can come to the rescue: a Darwinian control module, already possessing the competence for control in the adaptive contexts, can learn the

competence for control in a wide range of significantly novel contexts. In this section, I show that the kind of learning needed to make this reply is not the kind of learning MM can incorporate. The nativist control problem thus remains.

One preliminary: Both nativists and anti-nativists agree that through a lifetime of learning, we come to acquire a large number of psychological traits, and that innate information contributes to these learning outcomes. However, they disagree about "the character of the psychological systems that underlie the acquisition of psychological traits" (Margolis & Laurence, 2012, p. 3). MM theorists believe that the contribution of innate domain-specific information is more significant than that of experience in acquiring a particular trait.

We can understand this commitment better by looking into one of the central arguments that MM theorists utilize—namely, the poverty of the stimulus argument (PoSA) (Samuels, 2002; Tooby & Cosmides, 2016). The PoSA supports nativism for a particular psychological trait by establishing the following features in the learning mechanism responsible for its acquisition.

(1) The learning mechanism contains substantial innate domain-specific information.

The PoSA usually establishes this by showing that, based on observation, information in a learner's environment is inadequate to account for the acquired trait, or even if adequate, the subject has insufficient time to acquire the necessary information to develop the trait (Margolis & Laurence, 2012). Because the input is inadequate, what is missing has to be made up somewhere—potentially, by innate information built into the Darwinian learning modules. But PoSA does not merely establish that what is learned goes beyond what is experienced. It also shows that,

(2) Some of the innate domain-specific information acts as a strong bias or constraint in the learning process (e.g., in hypothesis generation).

This feature is established by showing that what is learned surpasses what is experienced in a way that cannot be accounted for without these innate biases and constraints. This is because "… the correct hypotheses are not at all the most natural ones for an unbiased learner…. Indeed, there are numerous alternatives that would be more natural to such a learner but that would lead the learner astray" (Margolis & Laurence, 2012). Finally, the abundant innate domain-specific information suggests that,

(3) This mechanism has a relatively constrained learning capacity.

Now, it is clear why the PoSA supports nativism—it gives innate domain-specific constraints a much more substantial role over experience in the acquisition of the particular trait. So, while both nativists and anti-nativists take both innate constraints and experience as determinants of a learning outcome, i.e., the acquired trait, only nativists believe that innate constraints, rather than experience, are the more significant determinants of the learning outcome.

We can illustrate this interplay between innate constraints and experience with the development landscape (Figure 3.3): the downwardly slanted surface represents all developmental possibilities and the ball's rolling trajectory represents an actual developmental course (i.e., the process of acquiring a particular trait). Certain developmental pathways (represented by the channels) are established by strong innate constraints such that it is very difficult for the developmental trajectory to depart from them. For example, according to linguistic nativists, this illustrates why all children end up learning their mother tongues reliably despite the impoverished, even misleading, guidance from environmental inputs.

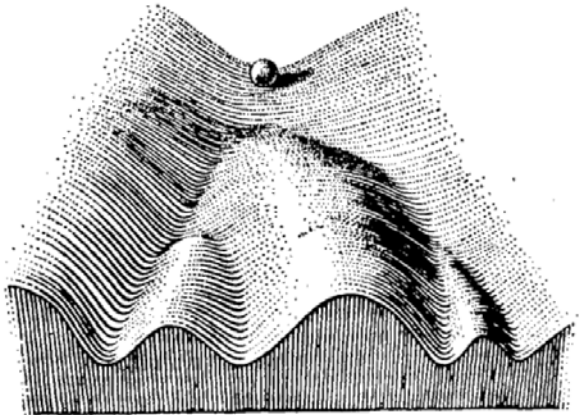

Figure 3.3 The development landscape. Excerpted from (Waddington, 1942).

Now, my second argument can be laid out as such:

(1) MM is committed to nativist learning, according to which innate domain-specific constraints are stronger determinants than experience in driving developmental outcomes.

(2) These innate domain-specific constraints will likely prevent the acquisition of competence for control for a significantly novel context.

(C) Nativist learning cannot enable Darwinian modules to perform control satisficingly for a wide range of significantly novel contexts—MM still faces the nativist information control problem.

The first premise has been established above, so let me focus on establishing the second. The innate domain-specific constraints are built-in by evolution because they have facilitated the acquisition of certain domain-specific competences that were useful in the EEA. In our case, these constraints would function to help acquire the competences for control for adaptive contexts. As a result, these strong innate constraints would create pathways in the development landscape that guide the learning trajectories toward these desired results.

However, the flip side of this is that these strong constraints will also form strong boundaries that impede the learning trajectories from diverging from the pathways. Because the competence for control for a significantly novel context is very different from that for an adaptive context, strong innate constraints designed for the latter will likely (though not necessarily) prevent the acquisition of the former.

For example, the control competence required to determine social exchange situations in the modern world is likely to be very different from one required in the EEA; so is the control competence required to route information from the object-recognition module to the naming module for the novel task of reading. As a result, nativist learning is likely to prevent the acquisition of the necessary control competences in both cases.

Let me illustrate the truth of the second premise with Figure 3.4. Intuitively, we can think of all possible learning outcomes for control as organized in a multi-dimensional trait space (represented in reduced two-dimensional space in Figure 3.4A). Experiences direct the learning trajectory, while innate constraints constitute sloped boundaries in surrounding areas in the space (the white circle). On the one hand, the innate constraints will facilitate the learning trajectory (the white arrow) to travel from the initial state (the triangle) to the acquired competence for an adaptive context (white star). On the other hand, the innate constraints will discourage the alternative learning trajectory (the gray arrow) from reaching competence for a novel context (the gray star). As adaptive and significantly novel contexts have very different information structures, the competence required for

one is very different from that required for the other, i.e., the two competences should be located relatively far away from each other in the trait space (Figure 3.4B). As a result, innate constraints that facilitate the acquisition of one are likely to discourage the acquisition of the other. Importantly, nativism requires the innate constraints to form a strong boundary, surrounding a relatively small area and impeding the learning trajectory from going beyond it.

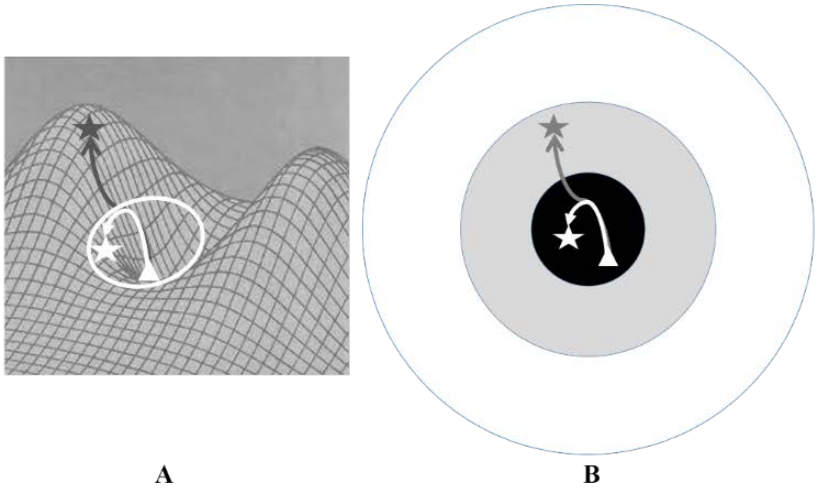

Figure 3.4 A. Trait space arranged in two-dimensional space, distorted by innate constraints. B. The trait space mapped onto our intuitive conception of flexibility.

Note that even if nativism comes in different strengths and the area constrained by the innate information may be larger for a weaker version of nativism, there is an ineliminable tension between the nativist commitment and the range of novel competences that are learnable. In other words, MM can incorporate as strong a learning process as they want; however, its nativist commitment will limit the strength of learning relative to the innate constraints. Claiming that MM's learning mechanisms can acquire the competences necessary for a wide range of significantly novel contexts is equivalent to denying the existence of strong innate constraints in learning and therefore, equivalent to denying nativism. In fact, this is a conclusion that simply follows conceptually from a reasonable conception of nativism. MM theorists do not see this because they take control for granted and fail to recognize the robust learning necessary to acquire the competence for control in a wide range of significantly novel contexts.

I've shown that nativist learning cannot help with the nativist information control problem. Darwinian control modules cannot acquire the competence to perform the control necessary for flexibility. MM theorists only seem to explain flexibility because they forget about their nativist commitment when it comes to the control modules.

5. Objections to the Nativist Information Control Problem

In this section, I address three potential objections to my arguments. As we will see, all of the objections aim to save MM by eliminating the gap between the range of novel contexts in which humans can perform satisficingly and the range of novel contexts in which Darwinian control modules can (learn to) perform satisficingly. However, I will show that none of the moves they make are acceptable; they either rely on problematic assumptions, lead to unacceptable consequences, or simply do not work.

5.1. Interactive Control Mechanisms Are More Flexible

MM theorists can object that my treatment of interactive control mechanisms as Darwinian modules is problematic. This objection asserts that interactive control mechanisms, due to the interactions among modules, are more flexible than I give them credit for. Although MM theorists never provide an explicit, mechanistic account for the required flexibility, let me grant for the sake of the argument that interaction can indeed make control mechanisms more flexible.[60] In response, I shall expose a dilemma concerning the nature of the flexibility-generating interaction.

On the one hand, one may suppose that the interactions are purely biological (i.e., non-psychological) processes. For example, Sperber suggests that control (in Sperber's terminology, relevance determination) can be handled by noncognitive mechanisms, such as ones that adjust the blood flow to allocate energy to relevant modules "without computing expected relevance" (2007, p. 67). Because purely biological developmental plasticity excludes learning, such noncognitive mechanisms are compatible with MM's nativist commitment. However, while purely biological developmental plasticity can no doubt generate some flexibility, it is unclear how it could extend the range far beyond adaptive contexts. Purely biological plasticity is unlikely to generate a novel complex mechanism that has a good fit with specific novel contexts. On the contrary, it is more likely for a purely biological system to lose its function altogether when its environment goes through a significant change. For example, enzymatic systems (e.g., a cell) do not exhibit satisficing performance across a wide range of significantly novel contexts, as they lose their function in the absence of the homeostatic processes to maintain the stability of the cellular environment within the range they are designed for. The inability for purely biological developmental plasticity to cope with significantly novel contexts is perhaps why a learning capacity had evolved to begin with.

On the other hand, one may suppose the flexibility-generating interactions are just robust learning processes. If so, it is unproblematic that they can expand the range of novel contexts in which control mechanisms can perform satisficingly. In making this move, however, MM theorists have to give up their nativist commitment for the specific capacity of control. In general, it is unproblematic to be anti-nativist about some specific capacities while being nativist about others. Moreover, a nativist position about the human mind, in general, can incorporate some anti-nativism about a limited set of capacities. However, giving up nativism about control will lead to a highly fragile nativist position. If a wide range of significantly novel complex competences for control can be learned, it is not clear why one should hold onto nativism about other cognitive capacities. Finally, control is too important a capacity to be anti-nativist about for the MM theorists—nativism about control is exactly what distinguishes them from Fodor, who believes in a central control system capable of robust learning.

In short, interactive control mechanisms may indeed be more flexible than I give them credit for. If so, however, such flexibility of control is either generated by purely biological plasticity and not sufficient for the sort of flexibility exhibited by humans, or generated by robust learning which results in a problematic nativist position.

5.2. The Intuitive Conception of Behavioral Flexibility is False

MM theorists could argue that the nativist information control problem is based on a false assumption. They might claim that humans do not possess the intuitive conception of behavioral

[60] See (Hung, 2014; Samuels, 2012) for criticisms on the interactive control model.

flexibility—the competences to perform tasks satisficingly *in a wide range of significantly different novel contexts*—but only possess the pessimistic conception of behavioral flexibility—the competences to perform tasks satisficingly *in a small range of novel contexts that are similar to adaptive ones.* The satisficing success enjoyed by humans on significantly novel contexts may, in turn, be explained by something like niche construction. Humans do not always passively adapt to the environments encountered; they also actively select environments and alter them to be more hospitable (Richerson & Boyd, 2005). So, the apparent high-level of behavioral flexibility humans exhibit may result from selecting and engineering novel contexts into ones more similar to adaptive ones. If this is granted, MM can adequately explain flexibility, since nativist learning is sufficient to acquire the control competences necessary for the pessimistic conception while another mechanism, such as niche construction, does the rest.

To reply, I shall stress that the intuitive conception of behavioral flexibility is a reasonable *empirical* assumption. It is supported by our record of developing diverse subsistence and social practices to survive in habitats ranging from the high Arctic to the balmy California coast (Boyd & Richerson, 2007). Moreover, consider the following empirical fact which appears to indicate diversity in human problem-solving: the information structures of the cheater detection task in the Pleistocene and in modern society are significantly different (Sterelny, 2012, 2014). Cheater detection is an important task as the existence of free-riders destabilizes the practice of social cooperation. Because Pleistocene humans foraged and lived together in relatively small groups, they had extensive knowledge of each other through direct observation. The expectation of repeated interaction presumably lowers the motivation to free-ride. However, with increased group size and specialization of labor since the Holocene, the human social world has become increasingly less informationally transparent and less intimate. Different types of information have become relevant to competent social exchange. An expanded folk psychology is now needed to assess the capacities and intentions of social partners for joint planning and action. Information about the partial biography and social assessment (e.g., reputation) of others has become necessary for tracking potential alliances as defection has become increasingly common. Therefore, the information structures for cheater detection have gone through significant changes. What is true for the cheater detection task is also true for a wide range of other tasks.[61]

In light of these considerations, the intuitive conception of behavioral flexibility is an empirically-supported claim that needs to be rebutted with empirical evidence—the burden of proof is on MM theorists to come up with independent empirical support for the pessimistic conception. That is, they need to show through detailed, case-by-case analysis of human problem-solving, evidence that the information structures of most of the novel contexts in which humans are capable of satisficing performance are similar to the information structures of adaptive contexts. At the very least, my arguments can be seen as conditional ones: if the intuitive conception of behavioral flexibility is true, MM fails to explain human intelligence due to its nativist commitment. MM cannot consistently endorse nativism while endorsing the intuitive conception of behavioral flexibility, contrary to what is often suggested by the writings of some of its proponents (Barrett, 2005b; Barrett & Kurzban, 2006; D. Sperber & Hirschfeld, 2006).

For example, when we look at one of Sperber's arguments carefully, we can see that it is only sound if we replace the explanandum with the pessimistic conception (1994). Sperber argues:

---

[61] See (Potts, 1997; Richerson & Boyd, 2012; Sterelny, 2006, 2012) for more discussion on the substantial difference between many contemporary and adaptive contexts in terms of information structure.

(1) Novel inputs (e.g., information about the alphabet) can be part of the "actual input domains" of some modules, even if they are not part of the "proper input domains."

The proper input domain refers to the type of input a module is designed to process; the actual input domain refers to the type of input it in fact processes.

(2) Modules are blind to the distinction of actual/proper input domains and will process inputs from either of them, as long as they meet the input criteria of a module (e.g., object recognition modules may be recruited to process words and sentences).

(C) MM can explain behavioral flexibility (e.g., the novel problem of reading).

Now, it should be obvious that this argument is not even valid. A crucial premise is missing:

(3) The modules that take novel inputs as their actual input domains also have the relevant competence to process them satisficingly.

For example, to play chess successfully requires more than having some module capable of taking information about chess as inputs, because this does not guarantee that the module will compute any satisficing moves of chess. What is required is that the module has the competence for chess. That is, Sperber fails to distinguish (a) the ability to take some novel inputs with (b) the competence to process them satisficingly in the relevant context. As a result, he also implicitly assumes that the competences for processing novel inputs in the novel contexts are identical or very similar to those for adaptive contexts—i.e., the pessimistic conception of behavioral flexibility.[62] Again, the failure of MM theorists to distinguish (input-)domain-specificity and context-dependence of Darwinian modules lead to an unsound argument.

In sum, MM theorists need to either show empirically that we do not enjoy the intuitive conception of behavioral flexibility or give up their nativist commitment; they cannot consistently embrace both.

5.3. Behavioral Flexibility is Achieved Socially

Another way that MM theorists can respond to my arguments is to claim that the intuitive conception of behavioral flexibility is a real phenomenon, but it is achieved socially, not individually. Science is our best example of such a social achievement. Through collective scientific endeavor, human beings solve problems that individuals cannot. As a result, so the response would go, MM correctly explains flexibility at the level where it is in fact achieved—namely, at the social level (Carruthers, 2003; Pinker, 2005; D. Sperber, 2007).[63] For example, Sperber explains:

In collective intellectual endeavors that are pursued over generations, and in science in particular, greater context sensitivity and greater relevance can be achieved… the explanation of these achievements calls for a kind of epidemiology of representations that looks at the effect of *the causal chaining of individual cognitive processes across populations*… (2007, p. 68, emphasis mine)

We can construe Sperber's response as follows:

---

[62] Barrett and Kurzban (Barrett, 2005b; Barrett & Kurzban, 2006) also make a similar argument to Sperber's.

[63] These responses are originally raised against Fodor's argument for the impossibility of cognitive science. Dominic Murphy (2006) has argued against this line of objection; for more recent discussion, see (Chow, 2016; Fuller & Samuels, 2014).

(1) To perform satisficingly in a context, Darwinian modules *across different individuals* need to be assembled properly (i.e., Sperber's *the causal chaining*) so that the relevant modules are engaged at the right time in the information processing.

(2) To assemble the relevant modules properly *across different individuals*, control needs to be performed at (almost) each step of the information processing.

(3) Darwinian modules can (learn to) perform control satisficingly at each step of the information processing for a wide range of novel contexts.

(C) *Social* MM models can perform satisficingly in a wide range of novel contexts, i.e., MM can explain flexibility, *achieved socially*.

This argument is similar to the implicit argument in MM's individualist explanation of behavioral flexibility discussed in Section 4.2., with the premises and conclusion slightly altered (the alterations are marked with italic). We can see immediately that the social explanation faces the same problem as the individualist explanation. Namely, premise (3) is false due to the nativist information control problem. Therefore, the nativist information control problem remains a threat to MM, even if the intuitive conception of behavioral flexibility is achieved socially.[64]

All of the three objections demonstrate to us again the ineliminable tension between nativism and flexibility. The achievement of collective scientific endeavors at the social level partly depends on a high degree of flexibility at the individual level, which in turn relies on (some) very flexible components at the neural-mechanistic level.

6. Conclusion and Implications for Control and Cognitive Architecture

Can a massively modular architecture explain human intelligence? I've argued for an answer in the negative. I have shown that as long as the nativist commitment remains a constituent of the MM hypothesis, MM models cannot overcome the information control problem to achieve flexibility. This remains so despite the incorporation of flexibility-enhancing features, such as self-organization, development, and learning. In short, to have a flexible self-organized whole, one cannot start with all inflexible parts—the consequence of MM's commitments to both heuristics and nativism. MM theorists fail to appreciate this because they have taken for granted the competence required for flexible control and assumed unintentionally an anti-nativist position on control. The failure to clearly distinguish domain-specificity and context-dependency may have contributed to this oversight. That is, MM fails to address the problem of intelligence with regard to the control problem: advocates of MM need to explain how intelligent control decisions can emerge from the interaction of Darwinian modules without violating their nativist commitment.

Finally, I take it to be a strength that my arguments do not depend on overly restrictive architectural commitments and epistemic requirements usually assigned to MM by their critics—my arguments remain sound despite granting MM resources many disallow. To conclude, I would like to discuss the broader implications of my arguments for accounts of cognitive architecture and control.

---

[64] I ignore the possibility that interpersonal information exchanges will completely take the burden of novel information routing away from control modules, and reduce control tasks within individual brains to contexts that are similar to adaptive ones. This possibility, if substantiated (which has not been done by MM theorists), may provide a viable objection to my arguments.

6.1. The Library Model of Cognition

My arguments can be extended to other cognitive architectures beyond MM. In the following, I briefly discuss how they can be modified to argue against what Samuels (1998a) calls that Library Model of Cognition (LMC), a very dominant view held among cognitive scientists. According to LMC, the human mind contains domain-general central mechanisms in addition to peripheral modules. Nevertheless, LMC differs from the Fodorian architecture in one important way: the domain-general central mechanisms, unlike central systems, do not perform optimal non-demonstrative inference. Instead, they contain many different innate domain-specific knowledge databases (hence, the "library" metaphor). LMC, as formulated, is a relatively weak thesis because Samuels refrains from making any theoretical commitments unnecessary for his purpose of accommodating some of MM's positive arguments. Here, I refer to a common and stronger version—what I will call "strong LMC"—which is also committed to nativism.

Strong LMC faces an information control problem very similar to the one that confronts MM. This is because, in order to exhibit flexibility, some control processes or algorithms (supposedly implemented by the domain-general mechanisms, instead of the modules) will have to choose which domain-specific knowledge to apply. Due to its nativist commitment, strong LMC cannot learn the required control competences. This perhaps is not surprising at all, since strong LMC only differs from MM in terms of their architectural features (one contains domain-general mechanisms, one does not). Yet, described in terms of information processing algorithms, both implement heuristics. Since the nativist control problem, at its core, does not depend specifically on the architecture of MM, it can be modified easily as an argument against strong LMC.

6.2. Lessons for Flexible Control Mechanisms

The main goal of this and the previous chapter is to argue against MM. Seen this way, the project is largely a negative one. However, engaging in this debate also yields positive lessons for what flexible control mechanisms require.

1. **Heuristics.** If Fodor is (at least roughly) correct that optimal non-demonstrative inference is intractable, control mechanisms need to implement heuristics.

2. **Anti-nativism.** As the commitment to heuristics is not optional, control mechanisms must be capable of robust learning that can reliably acquire complex competence for wide ranges of novel contexts. MM theorists have criticized the unreliability of domain-general learning (Tooby & Cosmides, 1995b). However, such criticism is misguided because learning need not rely on constraints that are available only intracranially. Niche construction and scaffolded learning can provide the necessary additional constraints to make domain-general learning both flexible and reliable (Menary, 2014).

3. **Distributed, self-organized control.** MM theorists have formulated the bulletin board architecture as a flexible type of control mechanism. The bulletin architecture can route a large amount of information, but only compute a small subset of the information using heuristics. Such soft-assembled interactive control mechanisms seem empirically plausible and should be further developed (Shanahan, 2012).

4. **Mechanism.** We need mechanistic accounts of learning and control.[65] Without mechanistic details, it is difficult to evaluate the extent to which an account succeeds or fails. Looking into the neuroscientific literature will prove useful in providing such

[65] I take a mechanistic account of Phenomenon X to be an explanatory model that explains X in terms of how it is produced by the causal interactions of spatially-structured entities with particular properties (Craver & Tabery, 2017).

details. It turns out that there are neural structures that meet these desirable features of flexible control mechanisms (e.g., the basal ganglia) (Hélie, Ell, & Ashby, 2015; Redgrave, Prescott, & Gurney, 1999).

Understanding how the cognitive system solves the control problem is essential to our explanatory endeavor in understanding human intelligence. It is in fact tantamount to the project of banishing the control homunculus hidden in the cognitive architecture. We've come a long way from the solution of the Cartesian intellect. It seems like, given what we learned from this debate and the available empirical theories, control mechanisms will look nothing like our folk psychological image of a neural commander-in-chief. How grotesque but well-designed such control mechanisms are, however, will be the topic for Chapter 6.

Part II

Society of Mind in the Twenty-First Century

4

Society of Mind in the Twenty-First Century I: The Hierarchical Embodied Cooperative Architecture

1. Introduction

In the last two chapters, I discussed two problematic solutions to the control problem that have been offered, respectively, by two dominant paradigms of the classical cognitive science and Evolutionary Psychology. In this chapter and Chapter 5, I will turn to the embodied cognitive science approach and evaluate its solution. I will first construct a novel embodied cognitive architecture, the hierarchical embodied cooperative architecture (HECA), through updating and revising the society of mind account of Daniel Dennett with current literature in the cognitive neurosciences. Then, I will evaluate HECA as an embodied cognitive science solution to the control problem.

The society of mind account (Dennett, 1991; Minsky, 1986) was first proposed over thirty years ago. Since then, it has been an important and influential alternative to classical cognitive architecture, and has paved the way for the more recent boom of embodied, extended, embedded, and enactive (4E) approaches to cognition (Chemero, 2009; Clark, 1998; Menary, 2010; Varela, Thompson, & Rosch, 1992). However, it is not clear whether, or to what extent, contemporary empirical research in cognitive neuroscience vindicates, develops, or rejects this account. Nor is it clear whether an empirically-updated society of mind account can make explanatory progress on issues of consciousness, intentionality, and human intelligence. More specifically, in our current context, we need to assess whether or not an empirically-updated society of mind account provides a better solution to the control problem.

In this chapter and the next, I will show that several large-scope contemporary empirical theories support the society of mind account.[66] These empirical theories include the *affordance competition hypothesis* for embodied motor decision-making (Cisek, 2012a; Cisek & Kalaska, 2010); *the hierarchical models of perception and action-control* (Botvinick, 2008, 2012; Felleman & van Essen, 1991; Fuster, 2004; Koechlin, Ody, & Kouneiher, 2003); *model-based and model-free reinforcement learning and control* (Daw & Dayan, 2014; Dayan & Niv, 2008), *the predictive mind* (Clark, 2013; Hohwy, 2013); *dual-process theories* (J. S. B. T. Evans, 2008; Frankish, 2010); and *the sequential sampling models* in decision-neuroscience (otherwise known as neuroeconomics) (Busemeyer & Johnson, 2004; Gold & Shadlen, 2007; Ratcliff & McKoon, 2008; Usher & McClelland, 2001; Wang, 2012). Despite significant differences in their target empirical phenomena and theoretical details, these theories nevertheless provide new computational and mechanistic (algorithmic or implementational) details (Bechtel, 2008) to some of the society of mind's core features (including distributed control and a simple message-passing strategy discussed in Chapter 1). Additionally, these empirical theories also develop and revise some of the society of mind account's important features. Specifically, I will demonstrate that contemporary empirical research provides three new insights into the society of mind account. Research shows, first of all, that the society of mind is hierarchically structured to mirror the causal hierarchy of the world. Second, it vindicates an embodied cognitive science interpretation of the society of mind account. Third, empirical research supplies the computational theory and mechanisms that govern the society of mind's dynamical "cooperative" decision-making processes.

In light of these three empirical developments, I will call this updated society of mind account the "hierarchical embodied cooperative architecture" (HECA). According to HECA, the human cognitive system consists of a large number of diverse cognitive mechanisms, which are located at different levels of various hierarchies in the brain. Different types of responses—including motor actions, perceptual decisions, motivational responses, and cognitive actions of various spatial and temporal scales—are specified and selected at mechanisms situated at different levels in the hierarchies. Also, these responses are specified and evaluated in parallel by embodied and autonomous information processes. When conflicting responses are generated, these embodied agents will cooperate with each other to determine the final actions to be executed. That is, adaptive actions are the result of the dynamical and cooperative interaction of embodied autonomous information processes.

This new architecture is "hierarchical" because the core competence involved in specifying and selecting a particular action emerges from the interaction of a hierarchically structured set of mechanisms, each of which possesses its own relevant competence. My architecture contrasts with the Fodorian cognitive architecture, where the core competence for action-selection is implemented singularly by the central systems (Fodor, 1983).

This new architecture is "embodied" because the autonomous information processes that specify and evaluate responses often (if not always) involve both sensory and motor mechanisms. These sensorimotor information processes are autonomous in the sense that their role in cognition is not dictated by other (higher-level) processes, but usurped through interactive processes.

It is a "cooperative" architecture because all decision-making is a collaborative process involving many autonomous information processes in the society of mind. Each decision is reached through

---

[66] By large-scope empirical theories, I mean theories that apply to several functional domains instead of a single domain, such as visual perception.

the accumulation of evidence for and against conflicting response options. This architecture again contrasts with the Fodorian architecture, where the action-selection is made by a single intelligent central system.

HECA further develops the society of mind architecture proposed by Daniel Dennett. However, it is significantly different from Dennett's account in a number of ways. First, HECA is narrower in scope, in the sense that I do not intend it to provide an account of consciousness, phenomenal or otherwise (Block, 1995). In other words, although it is compatible with the global workspace theory of consciousness (Baars, 1988; Dennett, 1991), it remains neutral as to whether consciousness can be explained in terms of global access to a mental representation. Second, HECA is wider in scope in the sense that it addresses the selection of responses or actions in general, including all internal and external ones. Dennett, in his development of the pandemonium model, tends to focus on overt verbal behavior or inner speech, even though he recognizes the wider application to other motor behaviors. In contrast, HECA explicitly applies to motor responses and cognitive, perceptual, and motivational responses, such as the control of visual attention, updating of working memory content, and sifting of the motivational drive. Third, HECA focuses on the cooperative, rather than the competitive, nature of the interaction emphasized by Dennett's pandemonium account. Working out the cooperative interaction as a computational theory will enable us to better understand how human intelligence can emerge from less intelligent neural mechanisms. Finally, HECA is a less metaphorical model and provides more computational and mechanistic details of information processes and their interaction, thanks to recent advances in empirical science mentioned above. As we will see in this and the following chapter, these new relevant details not only introduce new and distinctive features but will also illuminate new challenges for understanding the human cognitive system.

I will begin the next section with a review of Dennett's society of mind account, the *Pandemonium architecture*, and discuss its core features. Section 3 will introduce the novel features of HECA: (1) embodied information processes, (2) hierarchical structure and its related properties, and (3) dynamical cooperative decision-making and its epistemology. In the last section, I will highlight the distinctive features of HECA by comparing it with some dominant theories of the human mind. In the next chapter, we will look into recent empirical theories in cognitive neuroscience supporting HECA, as well as its remaining theoretical challenges with respect to the control problem.

2. A short historical primer on the society of mind account

In this section, I will introduce Dennett's Pandemonium architecture, which he developed on the basis of cognitive scientific findings in the 1970s, as a representative model of the Society of Mind account. I will focus on Pandemonium architecture's core features, particularly those relevant to my discussion of the control problem. Then I will show that the architecture shares embodied cognitive science's two core features: distributed control, and the simple message-passing strategy. This discussion will provide the background necessary to situate and highlight the distinctive features of my updated account: the Hierarchical Embodied Cooperative Architecture.

2.1. Dennett's Pandemonium Architecture

In *Consciousness Explained* (1991), Dennett introduces the Pandemonium architecture as a model of human cognitive systems. On the basis of this model, Dennett attempts an explanatory account of consciousness, intentionality, and linguistic behavior (among other things). The term "Pandemonium architecture" comes from Oliver Selfridge's pioneering model of pattern recognition (Selfridge, 1989). However, Dennett uses the term in a more generic sense to refer to all its direct and indirect descendant models, on which many '"demons" vied in parallel for

hegemony' (Dennett, 1991, p. 189). In the following, I will briefly introduce the Pandemonium architecture Dennett develops, focusing on aspects relevant to our current discussion of the control problem.

The first feature is Dennett's characterization of "demons." Dennett calls the functional information processing components in the Pandemonium architecture "demons" or "agents," following Marvin Minsky's usage in *The Society of Mind* (1986). A demon is a functionally-individuated information processing mechanism that has (1) a lineage, (2) a specialization, and, most importantly for our current discussion, (3) some degree of autonomy.

A demon has a lineage because it has an evolutionary history. Every demon forms part of our animal heritage and is evolved not for particularly modern tasks such as reading and writing, but ducking, predator-avoiding, berry-picking, and other essential tasks (Dennett, 1991, p. 254).

Moreover, a demon is specialized in its evolved function, but it is often recruited (often through learning) to perform new functions. It comes to perform its new specializations satisficingly, more or less, due to the combined effects of the innate endowment and powers of individual and/or social learning (Dennett, 1991, p. 254). Dennett does not address in details the computational theories and mechanistic accounts of these demons, aside from speculating that their development is achieved through an evolution-like process, "of generation-and-selection of patterns of neural activity in the cerebral cortex" (Dennett, 1991, p. 193). He believed, at the time of writing *Consciousness Explained*, that we can make better sense of these processes if we try to understand them at a "more general and abstract level" before "descending once again to the more mechanical level of the brain" (Dennett, 1991, p. 193).

Finally, a demon has some degree of autonomy. A demon is in no way autonomous in the sense a person is autonomous; however, it has more autonomy than a Fodorian "module." So, the contrast here is between what Dennett calls "bureaucratic models" (such as the Standard Account) and Dennett's own pandemonium models. On the one hand, the bureaucratic models posit dumb "bureaucratic" agents (Fodorian modules) that are incapable of learning and rely entirely on the control of the intelligent, learning-capable, central executives (the central systems) for the production of intelligent, flexible behaviors. This is because a Fodorian module's job description is strictly and carefully regulated by central systems and unmodifiable. That is, the intelligent central systems make "what to do" decisions first and then dictate the motor modules for processing the relevant "how to do" decisions (means/ends analysis for achieving the "what to do" decisions). For example, Fodorian motor modules are under the tight control of the central systems: they can only specify the motor movements of an action, e.g., walking across the street, the ultimate goal of which has been determined by the central systems. In other words, they cannot generate movements for an action the central system does not select or select against.

On the other hand, the pandemonium models posit autonomous agents, the demons. The control over what a demon can do is "usurped rather than delegated" (say, by central systems) and "in a process that is largely undesigned and opportunistic" (Dennett, 1991, p. 241). For examples, demons in the motor system often compete with each other for control over, and selection of, the overall goals of information processing of the motor system, such as walking across the street vs. staying at the pedestrian walk. Consequently, the demons determine their own ensuing job descriptions of which motor movements to be planned and specified.

This brings us to the second feature of the Pandemonium architecture, the process of achieving "hegemony": if cognitive systems consist of a collection of specialist brain circuits, the demons, how do cognitive systems generate intelligent behaviors? Again, without giving us a more detailed

computational or mechanistic account, Dennett suggests that it is through processes analogous to political processes in a society. Firstly, there is a redundancy of demons: for any specific function or response, there are many demons who specialize in it. As a result, these demons form competing coalitions to determine who takes charge at a particular moment. However, these coalitions take charge one after another "without elevating any one of them to long-term dictatorial power" (Dennett, 1991, p. 228). According to Dennett, when the demons are united under a common cause in this way, their information processing power is vastly enhanced. Of course, nothing guarantees that the winning coalition will be the one that is able to produce the best responses. Rather, Dennett acknowledges that "there are obviously at least as many bad ways for these conflicts to be resolved as good ways" (Dennett, 1991, p. 222). However, the transitions form coherent, purposeful sequences rather than spiraling into a chaotic anarchy. This is partly due to the good "habits" or information processing competence these demons acquire through a combination of innate constraints, individual learning and social learning. Some of these "habits" include what Dennett calls "self-stimulation," such as the subject's talking-to-oneself and diagraming-to-oneself, of which conscious reasoning is a special case.

A detailed example will help here. In illustrating how the Pandemonium architecture can generate satisficing linguistic behaviors that reflect a person's communicative intentions, Dennett proposes a process that is composed of a (redundant) society of contestants and judges. Numerous "word-demons" (which I will refer to as action-specifying agents) act as contestants who propose, in parallel, words or sentences to be expressed. Meanwhile numerous "content-demons" (which I will refer to as action-evaluating agents), act as judges who evaluate these words and sentences based on how well they satisfy the various criteria and constraints they are able to assess, including sounds, meanings, associations, and grammatical construction, etc. In this way, demons form collaborative and competing coalitions. Intentional and purposeful verbal behaviors, "emerge from a quasi-evolutionary processes [of words generating and selecting]… that involves the collaboration, partly serial, partly in parallel, of various subsystems none of which is capable on its own of performing—or ordering—a speech act" (Dennett, 1991, p. 239). In this pandemonium production process, the linguistic behavior and the communicative intention it embodies exist "as much an effect of the process as a cause"—they emerge as an effect of the processes, but once they emerge, they prompt a new wave of reactions from word-demons and shape and constrain the criteria content-demons use to evaluate them (Dennett, 1991, p. 240). In short:

> Instead of a determinate content in a particular functional place, waiting to be Englished by subroutines, there is a still-incompletely-determined mind-set distributed around in the brain and constraining a composition process which in the course of time can actually feed back to make adjustments or revisions, further determining the expressive task that set the composition process in motion in the first place. (Dennett, 1991, p. 241)

Despite Dennett's colorful and plausible metaphors, the foregoing picture resembles "a sort of internal political miracle" (Dennett, 1991, p. 228). Without a more detailed model, it is not clear how such complex, self-organized activities in a cognitive system can reliably result in more or less intelligent behaviors—this is an explanatory gap that this chapter and the next are intended to help fill.

I will now conclude this section with a final comment about some of the distinctive features of the pandemonium model developed in *Consciousness Explained*. Even though Dennett recognizes that there are multiple channels of parallel pandemonium competitions (Dennett, 1991, p. 253), he often focuses his discussion on some subset of them that he thinks are responsible for producing the "Joycean machine"—a virtual, serial information processing "program" installed on the parallel information processing "hardware" of human cognitive systems. First of all, he develops an

extensive account of linguistic behavior, not only because it is the hallmark of human intelligence, but also because it is a strictly serial behavior; language-users can only speak one sentence at a time. Although Dennett explicitly generalizes the same pandemonium processes to all "intentional action across the board" (Dennett, 1991, p. 251), he nevertheless fails to consider and apply them, at least explicitly, to a large set of phenomena, for example, unintentional actions, and a variety of cognitive actions, such as control of perceptual attention and the updating of working memory.

Relatedly, his discussion also concerns pandemonium competition that determines what information would gain an additional functional role through being broadcast to the whole cognitive system, and in doing so, according to Dennett, become conscious. Here, his unique development of the pandemonium account concerning consciousness is part of the global workspace theory of consciousness (Baars, 1988). Although HECA is compatible with Dennett's theory of consciousness, I remain agnostic about whether HECA or Dennett's theory could adequately explain consciousness or not.

2.2. The Society of Mind as an Embodied Cognitive Science Approach

In the last section, I reviewed Dennett's Pandemonium architecture as a representative model of the Society of Mind account. Now I will show that it is also representative of the embodied cognitive science approach to the control problem, because it possesses two core features: distributed control and the simple message-passing strategy.

Control is entirely distributed in the Pandemonium architecture. As we discussed earlier, there are no central controllers, such as the central systems, that have the final say or ultimate control of the architecture's overall behavior. Instead, responses (internal or external) are selected by autonomous action-specifying and action-evaluating agents distributed across the cognitive system. This is true even of the selection of linguistic behaviors or the selection of information to be broadcasted through the global workspace—what gets to be broadcasted or which agents get to take control of the linguistic outputs is not determined by a privileged central system, but by a hegemony formed by the winning coalition of agents.

Pandemonium architecture also utilizes a "simple message-passing strategy." Although neither this term nor the strategy it denotes is explicitly discussed by Dennett, the selection of response through power competition is a metaphor of a power struggle between competing coalitions of agents through the pushes and pulls of forces, not rational discourse. This metaphor can be best understood as based on the mutual inhibition and activation between neural mechanisms, rather than that of the exchange of reasons or other messages with rich content.

To conclude Section 2, we've seen Dennett's interesting development of the pandemonium model as an embodied cognitive science approach to cognitive architecture. As Dennett acknowledges himself, this model is described at a very general and abstract level. In fact, we can see this coarse-grained picture as a task-analysis (as part of the computational level of analysis) of behavior generation (Marr, 1982; Piccinini & Craver, 2011). That is, Dennett decomposes the computational task of behavior generation into the sub-tasks of: (1) demons' individual activities of action-specification (word demons) and action-evaluation (content demons), and (2) their coalition formation and competition, without describing much the computational theories and mechanistic details of either. His focus on high-level task analysis is valuable and certainly justified, as the relevant neuroscientific details were simply not available at the time. However, recent developments in cognitive science have provided us an abundancy of relevant information at various levels of brain organization. It is thus an opportune time to consult contemporary empirical theories in order to provide a more detail-rich cognitive architecture.

3. The Hierarchical Embodied Cooperative Architecture

In the rest of this chapter, I will develop an empirically-updated society of mind account, HECA, by building on Dennett's work. The result is an image of the mind that is robustly supported by the empirical literature and rich with new computational and mechanistic details. These new details will not only enable us to better understand how the human mind copes with the control problem and produces intelligent behavior as an embodied and situated system, but they will also expose new challenges that our cognitive system needs to overcome. These challenges will be discussed in the next chapter. In this section, I will first sketch out the rough shape of HECA with a thick brush. In the subsequent three sections, I will turn to three of HECA's developments. The first development concerns the embodied character of the autonomous agents or information processes in the architecture; the second concerns the hierarchical structure of the neural mechanisms; and the third concerns the cooperative nature of the interaction between neural agents in the production of intelligent behaviors.

The agents or "demons" in HECA are "information processes" that specify or evaluate action or response options—that is to say, they are action-specifying and action-evaluating information processes. Moreover, (at least many of) these information processes are embodied information processes: they are, at minimum, subserved by both perceptual and motor mechanisms, and can in some cases involve bodily or environmental mechanisms. Additionally, the responses or actions that are specified or evaluated include not just the external motor actions (e.g., the movement involved in reaching for a soda or walking toward the supermarket), but also internal actions, such as perceptual decisions (e.g., determining the motion of a moving car), cognitive actions (e.g., directing one's attention to an unexpected alarm), or motivational/emotional expressions (e.g., getting into fight or flight mode).

Moreover, these information processes are subserved by neural mechanisms structured in various hierarchies, e.g., various perceptual, motor, or motivational hierarchies (Figure 4.1). The neural mechanisms situated in a higher level in a hierarchy will receive information from a wider range of lower-level neural mechanisms. Also, higher-level mechanisms will process this wider range of lower-level information to extract information about the deeper structure of the world (including the cognitive system itself) and use such information for more complex information processing. For example, in the visual hierarchy, higher-level neural mechanisms in the polymodal association area may process information from different perceptual modalities about the functional identity of an object (say, a train), compared to the more surface-level information available at the lower-level mechanisms in the unimodal association areas (e.g., colors and shapes of the object).

On the other hand, higher-level mechanisms can provide top-down contextual information to lower-level mechanisms so that they can be sensitive to deeper features of the world. For example, the higher-level motor control mechanisms—say in the prefrontal cortex—can bias the selection of specific motor movements at the lower-level of the motor cortex (say, driving straight or turning right) by providing contextual information about the current higher-level goals of the agent (e.g., one is going to work or going to the gym).

As a result, not all information processes are, so to speak, "created equal." They differ systematically in several aspects of their information processing characteristics, such as flexibility, speed, and capacity. This is because they involve neural mechanisms at different levels in various information processing hierarchies. Specifically, there is a larger number of lower-level information processes; Relatively, there are a smaller number of higher-level information processes. Lower-level information processes are information processes subserved only by lower-level mechanisms. These include, for example, mechanisms that only involve the unimodal association area and the

primary sensory area in the perceptual hierarchy, as well as premotor and primary motor areas of the motor hierarchy (as illustrated in Figure 4.1). On the other hand, higher-level information processes are subserved by both higher-level and lower-level mechanisms, for example, mechanisms that involve all areas in the perceptual and the motor hierarchies (see Figure 4.1). Moreover, lower-level information processes are less flexible yet faster in speed, while higher-level information processes perform more flexible information processing and have slower processing speed.

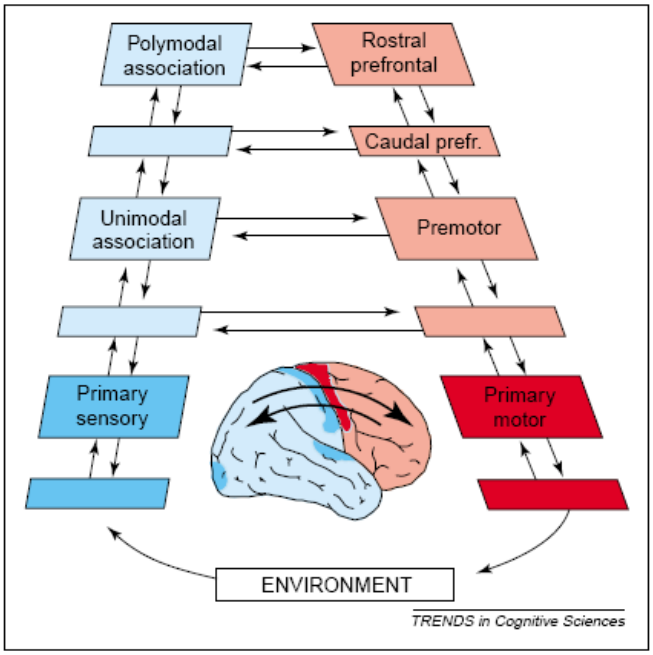

Figure 4.1 A schematic representation of an integrative perception-action hierarchy. The perceptual hierarchy is represented in blue, and the action-control hierarchy is represented in red. Hierarchies of mechanisms in different brain areas are arranged according to their forward and backward neuronal connections. There are also reciprocal connections between areas of the same rank within the hierarchies. It is only a schematic representation because the specific neural mechanisms within each sensory or motor area are not shown. Reprinted from Fuster (2004)

Also, information processes are autonomous in the same sense as Dennett's demons. That is, control over what a particular information process can do is not carefully delegated, but usurped through the interaction of multiple processes. In fact, no information processes are under strict control of the others, higher-level or otherwise. Instead, information processes of different levels work together to influence the overall operation of the cognitive system and in turn, determine what they can do in the mental economy. Specifically, lower-level Information processes can specify, support, or even successfully select, internal and external actions without the collaboration of higher-level information processes.

In HECA, response-selection is distributed across a large number of loci at all levels of many different hierarchies; different types of response options are specified in their associated neural mechanisms respectively. Moreover, during the period of decision-making, options are continuously specified and refined by action-specifying information processes, and they are continuously evaluated by action-evaluating information processes as well. The order and location of specification and selection for specific responses is important: the response options are specified first, and then selected in the same associated neural mechanism. There are interesting empirical implications, here: contrary to the bureaucratic model of the Standard Account discussed in Section 2, where decisions of "what to do" are (1) made in the central systems, and (2) *prior* to decisions of

"how to do" are made in the peripheral motor modules, HECA predicts a different order, similar to Paul Cisek's model (Cisek, 2012a) to be discussed in Section 2 of the next chapter. That is, the "what to do" decisions are made in the same brain area as "how to do" decisions, and occur either s*ubsequently* or at *the same time* as them.

Finally, actions are selected through collaborative epistemic processes. These processes are analogous to forms of collective decision-making, such as the majority vote, that are characteristic of democratic societies. Like Dennett's Pandemonium architecture, there is a redundancy of information processes and efforts: different types of information processes, specializing in different but perhaps overlapping aspects of the task, are involved in the selection of the same response. Specifically, simple (positive or negative) scalar evaluative signals for alternative response options, which are generated by relevant action-evaluating information processes, are accumulated dynamically (Figure 4.2). These evaluative signals are also called "*decision variables*" (DVs), as they are simple variables based on which decision is made. This accumulation of DVs will continue until one of the response options' accumulated DVs reaches the decision threshold**,** at which point the decision will be made— i.e., the response option with the highest accumulated DVs at the time is chosen, regardless of any possible future incoming evaluative signals. The selected response is more likely to be close to optimal than the other alternatives, because it reflects the collective judgment of different specialists of the "society of mind." The selected response at one location can then influence the response selected at various other loci, by participating in further action-specifying or action-evaluating processes. The final intelligent behaviors of an agent will reflect and emerge out of an overall (more or less) consensus reached at various mechanisms in the cognitive system.

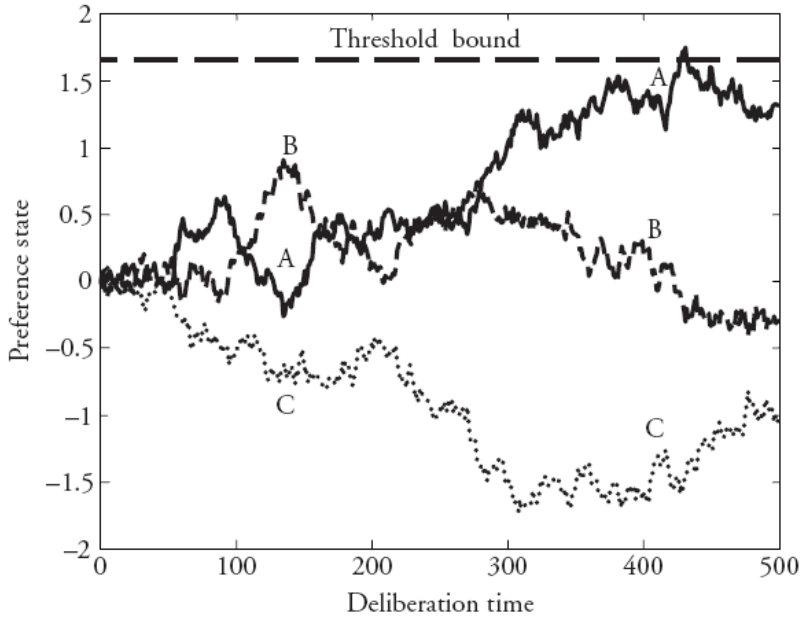

Figure 4.2 A selection process for a choice among three response options. The horizontal axis represents decision time; the trajectory represents the accumulated evaluative signals (DVs) for each option at each time point. The option A crosses the threshold and is chosen for execution around 400ms. Reprinted from (Busemeyer & Johnson, 2004).

Let me illustrate this process with one example. Imagine a very big family who has recently obtained a small fortune, and needs to decide collectively what to do with the money so that the family unit (taken as a whole) can benefit the most. Some members of the family propose a few options for using the money, then each member evaluates how each of the options will benefit the family as a whole. Different members have different perspectives on the best interests of the family,

so they will consider different features of the options and the family's current situation, and also weigh these features differently in their evaluation. Information about different options and current situations may also arrive, become available, or grow salient at different points in time due to factors internal or external to the family (e.g., the price of a certain goods increases, or debates in the family make certain considerations salient). This will influence how the family members take different pieces of information into consideration, and may lead to new options being proposed. Finally, family members also have different styles of evaluation: some prefer quick and dirty evaluation—making judgments based a small amount of information and simple, rule-of-thumb calculations— while others take a more careful approach, thinking through many pieces of relevant information and deliberating how each option will generate possible future outcomes and benefits. After evaluating each option, each member writes down his or her evaluation in a number that reflects his or her estimate of its value. Then, the evaluation for each option is aggregated, respectively, to reflect the collective judgment of each option's optimality. The option with the highest aggregated value is then chosen. However, because of the time pressure (e.g., perhaps the currency in their nation of residence is depreciated very quickly), the final decision cannot wait until all members evaluate all options—it would be too late to do anything with the money. As a result, a reasonable criterion (a threshold) is set to balance the accuracy of the final decision with the time spent on this collective deliberation. Once an option reaches the threshold (obtains an aggregated value higher than that), the family's decision will be made. This settled decision will then influence and constraint the downstream decisions the family make.

In the following three sections, I will elaborate the three major developments summarized above: the embodied autonomous information processes, the hierarchical structure of neural mechanisms and information processes, and the collaborative dynamical decision-making. I will leave the review of empirical theories until the next chapter, where I show that HECA is well supported by contemporary empirical models of cognition. In turn, these recent models supply important mechanistic details of to the information processes involved in this architecture.

3.1. The Embodied Agent Thesis

In this section, I will elaborate on one of the distinctive developments of my empirically-updated society of mind model, the Embodied Agent Thesis. This thesis further specifies the nature of Dennett's agents in the Pandemonium architecture. I will first state the Embodied Agent Thesis in its most concise form, before articulating a few different dimensions along which it can be interpreted. Having done this, I will then clarify HECA's exact commitment to embodiment.

The Embodied Agent Thesis can be stated concisely as follows,

> **Embodied Agent Thesis**: the action-specifying and action-evaluating autonomous information processes are embodied.

According to the embodied cognitive science approach, cognition is dynamical and representationally distributed, and depends heavily on not only neural, but also bodily and perhaps environmental mechanisms. I discussed in detail the broader embodied cognitive science paradigm in Chapter 1. Here, I will focus on clarifying the Embodied Agent Thesis.

Because embodied cognitive science is not a uniform research program, there are several dimensions upon which the Embodied Agent Thesis can be developed. The precise form in which the thesis is developed along these dimensions will, in turn, determine how "radical" HECA is. In the following, I briefly discuss three important dimensions:

Causal/constitutive dependence. The dependence of certain cognitive capacities on the body can be spelled out in either the causal or the constitutive sense. Constitutive dependence requires that the vehicles of these cognitive capacities include the (non-neural) body, an extremely counterintuitive position. Causal dependence, on the other hand, merely requires the recognition that the body is essential to bringing about (developmentally and/or causally) these cognitive capacities (Block, 2005).

Offline/online embodiment. The precise nature of somatic dependence can also be spelled out in two different ways. According to the thesis of online embodiment, cognitive processes depend on the organism's (non-neural) body for their realization; yet for the offline embodiment, cognitive processes merely depend on the sensorimotor parts of the nervous system for realization—a much weaker position (Robbins & Aydede, 2008).

Representational/anti-representational embodiment. Embodied cognition theorists differ in their opinions about the roles representations play in cognition. Self-identified 'radical' theorists believe that there is little to no role for any types of internal representations, while moderates believe that internal representations of some sort often play indispensable roles in at least some cognitive capacities, and are the key to human intelligence (Clark, 2001; Van Gelder, 1995).

Using the three dimensions discussed above, the Embodied Agent Thesis can be spelled out in more detail as the following three statements:

> First, all autonomous information processes depend constitutively on sensorimotor mechanisms (i.e., universal constitutive dependence on offline embodiment).
>
> Second, some autonomous information processes depend not just causally but constitutively on bodily or environmental mechanisms (i.e., existential constitutive dependence on online embodiment).
>
> Finally, many of the important information processes depend constitutively on internal representations—especially the modality-specific and action-centric ones (i.e., representational embodiment).

The Embodied Agent Thesis is a further development on the nature of demons in the Pandemonium architecture. In the next chapter, I will review empirical theories that support the Embodied Agent Thesis. I will also discuss, in the last section of this chapter, where this thesis places HECA on a spectrum of available positions ranging from "right wing" classical cognitive science to "left wing" radical embodied cognitive science.

3.2. The Hierarchical Structure Thesis

In this section, I clarify the second thesis of HECA. It further develops one of the key insights of the Pandemonium architecture: for any given task, there is a redundancy of agents, and there is a division of labor among them such that they specialize in different aspects of the same task. In the following, I will first introduce the Hierarchical Structure Thesis, then explain the key concepts involved. I end by articulating the significance of this thesis for the original Dennettian idea.

This thesis can be stated concisely as:

> **Hierarchical Structure Thesis:** neural mechanisms are structured hierarchically, and they represent the causal structures of the world (including the subject) at different depths

and/or process information with computation of different complexities; as a result, they subserve different levels of information processes that vary systematically in their flexibility, capacity, and speed.

First, let me clarify the concept of hierarchy, of which there are two senses in the empirical literature. According to the strong sense of hierarchy, an information-processing hierarchy is composed of a series of functionally-individuated mental mechanisms, arranged in the order of their forward and backward neuronal connections. Here, "forward" and "backward" connections are technical terms in neuro-anatomy, referring to types of neuronal projections that originate in specific layers of the cortex (Bechtel, 2008, p. 123). Generally, forward projections originate in the middle layers, backward projections in the lowest and highest layers, and lateral connections from all layers. Because of their distinct origins, we can use these as clues to arrange different functional areas into hierarchies, despite the prevalent reciprocal connections between them. Whether a particular mechanism is a higher-level or lower-level one is, as a result, only relative to the other mechanisms in the same hierarchy. Figure 4.1 (above), a hierarchical model of perception and action control, illustrates an information-processing hierarchy in this strong sense (Fuster, 2004).[67]

Hierarchy in the weak sense is, on the other hand, a less defined concept. Some models of cognition distinguish higher-level mechanisms from lower-level mechanisms without resorting to anatomical features. Instead, higher-level mechanisms seem to have certain superior, normatively-evaluable features, such as the capacity to produce a normatively correct answer to a question. In addition, they are also associated with a cluster of features such as slower speed and limited capacity (Evans & Frankish, 2009; Kahneman, 2011). For example, the dual-process theory (as illustrated in Figure 4.3) does not explicitly contend that the lower-level mechanisms involved in type 1 processes bear any particular anatomical relations with higher-level mechanisms involved in the type 2 process.

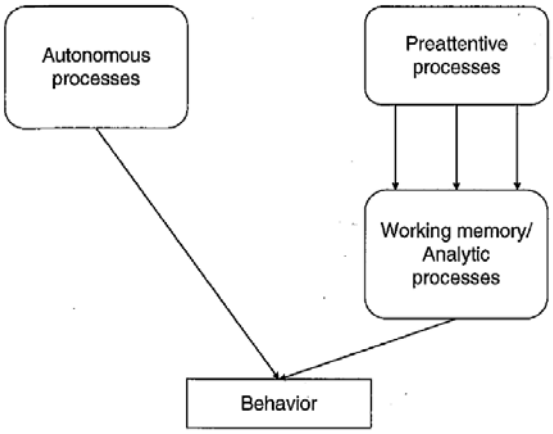

Figure 4.3 A representation of two forms of dual-process theory. The first form of dual-process theory, the parallel-competitive theory, contrasts autonomous processes (type 1 processes) with analytic processes (type 2 processes); the second form, the default-interventionist theory contrasts preattentive processes (type 1) processes with analytic processes (type 2 process). Reprinted from (J. S. B. T. Evans, 2009, p. 43)

[67] An information-processing hierarchy needs to be distinguished from a hierarchy of organization, or a level of organization. Level of organization refers to compositional levels, that is, "hierarchical division of stuff…organized by part-whole relations, in which wholes at one level function as parts at the next (and at all higher) levels" (Wimsatt, 1994, p. 222). Also, it is distinct from Marr's levels of understanding, which are different perspectives one could take when studying information-processing systems (Marr, 1982, Chapter 1).

In HECA, higher-level mechanisms play roles that are similar to what Clark calls "neural control structures." Neural control structures are "neural circuits, structures, or processes whose primary role is to modulate the activity of other neural circuits, structures, or processes" (Clark, 1998, p. 136). Instead of tracking and controlling external states of affairs, they track and control the inner economy, by modulating, for example, the flow and activity of other information processes. For example, neural control structures can modulate one's auditory attention so that the auditory processing is focused on a specific target rather than the other.

Like the neural control structures, higher-level mechanisms in HECA enable higher-level cognitive capacities. These capacities, which include selecting responses based on prospective planning (as opposed to habitual response-selection), and representing complex states of the world, are achieved by coordinating the activities of lower-level mechanisms using simple messages.[68] However, in contrast to Clark's characterization of neural control structures, higher-level mechanisms enable higher-level functions by playing important roles in tracking and controlling external states of affairs (such as representing deeper structures of the world and the agent) and/or directly manipulating information using more complex algorithms. This view of higher-level mechanisms is supported by the empirical theories reviewed in the next chapter, including one that Clark endorses.

A few comments before we proceed. First, I will assume the strong sense of hierarchy—particularly in relation to the analysis and theoretical consequences I will draw out in this chapter and the next. However, HECA need not commit to this construal, and most of the analysis will remain true regardless of whether mechanisms in the cognitive system are arranged in the strong or weak sense of hierarchy. However, I do believe that, empirically, the cluster of properties related to higher-level mechanisms in the weak sense are in part causally produced by the way mechanisms are arranged anatomically in the hierarchy in the strong sense. As a result, the fact that I have assumed the strong sense of hierarchy should be born in mind when the features of a cognitive hierarchy are under discussion.

Second, HECA does not assume that human cognitive systems have simple and orderly hierarchical structures. A cognitive system has a simple hierarchical structure if its mental mechanisms are organized into only a few hierarchies (e.g., only one perceptual and one motor hierarchies, instead of many different ones in the perceptual domain, motor domain, or motivational domain, etc.). A cognitive system has an orderly hierarchical structure if its hierarchies do not crisscross in such a way that a single mental mechanism can be placed at more than one of the hierarchies. Empirically, human cognitive systems may have neither simple nor orderly hierarchical structures; however, whichever way the empirical chips fall, should not affect the analysis I draw in this and the following chapters.[69]

Because the neural mechanisms are structured hierarchically, the selection of a given response—say, the motor action of kicking a soccer ball—can involve information processes of different levels. For example, the motor actions of kicking a soccer ball during a game can be generated by

---

[68] However, see the predictive mind model discussed in the next chapter for a reversed view where the lower-level mechanisms modulate the higher-level ones with simple predictive error messages (Clark, 2013; Hohwy, 2013). As we will see, this does not affect the general outlook of HECA.

[69] When two mechanisms are both involved in several different hierarchies, it may not always be possible to distinguish which is the higher-level one. Moreover, the dualistic distinction of higher-level and lower-level processes is merely a convenient way of referring to processes that can be arranged on a scale of many levels. Neither of the two facts should be consequential to the core claims in this and the following chapters.

information processes subserved by lower-level mechanisms for movement generation together with higher-level mechanisms for deliberate planning. On the other hand, the same action can also be generated by lower-level sensorimotor processes, e.g., a more reflex-like reaction when a soccer ball shows up unexpectedly from the player's left side. The information processes generating the first action are higher-level information processes because they involve higher-level mechanisms in addition to lower-level ones; the information processes generating the second action are, in contrast, lower-level information processes because they only involve lower-level mechanisms.

Distinguishing information processing that involves a particular higher-level mechanism from that which involves only lower-level mechanisms can be an intricate task. For example, it may be hard to distinguish two of the processes in Figure 4.1(p.66), one involving mechanisms in the polymodal association areas and the rostral prefrontal cortex in the top level vs. one involving only mechanisms in the unimodal association areas and premotor areas at the lower level. This is because the information processing at any particular level is constantly under modulation from its higher-level mechanisms through backward projections. In addition, this modulation effect is cumulative, so that, for example, even the premotor area at the lower level of the hierarchy is modulated (indirectly) by the rostral LPFC area at the top level (Fuster, 2004, p. 144; Koechlin et al., 2003). Nonetheless, we can draw a meaningful distinction between (a) information processes that "truly involve" a particular higher-level mechanism and (b) ones that are "merely modulated" by the higher-level mechanism. The difference is that, in the former case, the modulation from higher-level mechanisms is causally sensitive to the information processing at its lower levels, i.e., the modulation would be different if the information processing were different in some relevant ways. This is because the modulation in the former case results from the higher-level mechanisms' information processing of the relevant information coming from the lower level mechanisms. However, in the latter case, the modulation is not causally sensitive to the information processing at its lower levels, and would remain the same even if the lower-level information processing were different in some relevant way.

To return to the soccer example: in a higher-level information process, the modulation from higher-level mechanisms in the prefrontal cortex is a function of the information processing in the lower-level mechanisms, such that if the soccer ball came from a different direction with a different speed, the modulation would be different. However, in a lower-level information process, the modulation from higher-level mechanisms would remain the same, even if the lower-level information processing were different due to the different direction and speed the soccer ball approached with.

So far, we have discussed how the same type of response-selection in HECA can involve different levels of action-specification and action-selection information processes, which in turn involve different levels of neural mechanisms. Because of the involvement of different levels of mechanisms, higher-level and lower-level information processes also differ systematically in their flexibility, speed, and capacity. In the following, let me clarify the concepts of flexibility, speed, and capacity.

Flexibility

Information processes in HECA are not equal with regard to their information processing flexibility: Higher-level information processes are more flexible compared to lower-level information processes. Flexibility is the capacity of an information process to produce optimal responses in a wide range of contexts. An information process is more flexible than another if it can produce more optimal responses in a wider range of contexts. That is, an information processes' level of flexibility reflects its capacity to produce responses (1) of different degrees of optimality, and do so (2) in different ranges of contexts.

As discussed in Chapter 2, "context" refers to the information structure of a particular task in a particular environment, which includes relevant factors that are external to the cognitive system in evaluation—for example, features of the agent's body and the environment they inhabit. The most optimal solution in a context is one whose consequence have the highest utility of all plausible responses in that particular context.[70] Hence, what response is optimal for an information process is a function of its current context. What counts as the optimal route from the central station to the university, for example, depends on whether it is raining (the external environment), whether the agent has recently strained his or her ankle (bodily condition), and whether the agent is in a hurry (the specific constraints of the task).

A more flexible action-specifying information process can specify more optimal response options in a wider range of contexts. An action-evaluating information process with greater flexibility can generate more optimal DVs (i.e., DVs that are accurate) in a wider range of contexts. For an information process to be flexible, it needs to be sensitive to the goals in the particular contexts and the relevant particularity of the environments. That is, it will need to gather a larger volume of information about the contexts and compute and extract the relevant details to inform the generation of responses. Higher-level information processes are more flexible because they involve higher-level mechanisms that are capable of representing deeper structures of the world, and/or more complex information processing, to meet the demand of the complexity of more contexts.

Finally, we need to distinguish flexibility and reliability. I define reliability as the likelihood of an information process to generate optimal response options or accurate evaluations in *a particular context* under consideration. In other words, flexibility is the capacity to be reliable across contexts. As a result, it is possible for an information process to be highly reliable (in a particular context under consideration) but not flexible at all. Also, the fact that higher-level information processes are more flexible than the lower-level ones does not entail that they will always produce more optimal responses. In any particular instance, it is possible that a lower-level information process will happen to specify a more optimal response, or generate a more accurate DV, than a higher-level one does. It is especially likely if the lower-level process is specialized (through evolution or learning) in that particular context. However, when we look at the general trend in a wide range of contexts, higher-level processes will tend to produce more optimal responses. That is, flexibility reflects optimality across contexts, not just within a particular context.

In conclusion, higher-level information processes are more flexible compared to lower-level ones. Higher-level action-specification processes are more likely to specify optimal responses in a wide range of contexts, and higher-level action-evaluating processes are more likely to provide accurate DVs in a wide range of contexts.

Speed

In HECA, different levels of information processes also operate at different speeds. For a particular response-selection, the higher-level information processes involved are slower compared to the lower-level processes involved. Here, speed is a function of reaction time, the time elapsing between the relevant input and output. For example, for a perceptual decision task, the reaction time may be between the presentation of a perceptual stimulus (a triangle on the right side of the visual field) and the production of relevant neuronal responses (the activation in the motor area

[70] Which consequence has the highest utility is relative to different standpoints of utility assessment, e.g., subjective preference, objective value, individual fitness, group fitness, etc. I will remain neutral as to what type of utility is relevant for now.

responsible for pushing the right button). Higher-level information processes are slower because they take more time to specify their response options (of pushing the right or left button) and/or generate DVs (for or against either option).

This feature should not appear controversial, given: (1) it takes a longer time for information to travel up and down the hierarchy; (2) higher-level mechanisms partly depend on lower-level mechanisms for inputs; and (3) more complex information processing often correlates with longer processing time. This may explain why some computational modeling literature does not even bother with references when making this claim (Wiecki & Frank, 2013).

In short, higher-level information processes operate at a lower speed compared to the lower-level processes involved in the same response-selection task. It takes more time for them to specify relevant responses and to provide DVs for them.

Capacity

Another important feature of HECA is that there is, in total, a more limited capacity of higher-level information processing. That is, at any given moment, there are fewer higher-level information processes than lower-level ones across different response-selection tasks. As a result, only a subset of the total response-selection tasks can involve higher-level processes; they are a more limited resource that needs to be carefully managed so that they only engage in response-selections that will benefit significantly from their involvement.

One reason higher-level information processing is capacity-limited is suggested by Jonathan Cohen (Cohen, Dunbar, & McClelland, 1990, p. 357): in cognitive systems, several lower-level mechanisms often rely on the same higher-level mechanism to process their information. When two or more sources of information need to be processed by the same higher-level mechanism in disparate ways, the higher-level mechanism will not be able to support the processing of all of them at the same time. As a result, only some of them can gain access to the higher-level mechanism for further processing. In this sense, higher-level information processing can be thought of as capacity-limited. In short, there is more limited capacity for higher-level information processes than for lower-level ones. As a result, higher-level resources need to be managed carefully to ensure that they are engaged only selectively.

The Hierarchical Structure Thesis situates HECA midway between the "bureaucratic models" of the Standard Account and Dennett's pandemonium models discussed in Section 2. On the one hand, HECA is not a bureaucratic model because it does not posit dumb "bureaucratic" agents that are incapable of learning and who operate under the strict control of central systems for the production of intelligent, flexible behaviors. On the other hand, my model is unlike Dennett's pandemonium model, insofar as higher-level and lower-level mechanisms are subject to a greater division of labor. The higher-level mechanisms, specialized in tracking deeper structures of the world and in more complex computation, can modulate and constrain the activities of the lower-level mechanisms, even if they do not dictate them.[71] That is, HECA is distinct from pandemonium models that posit absolutely no hierarchy of control structure, but only "lots of duplication of effort, waste motion, interference, periods of chaos, and layabouts with no fixed job description" (Dennett, 1991, p. 261).

---

[71] Examples of higher-level mechanisms modulating and constraining the activity of lower-level mechanisms can be seen from the discussion of supporting empirical theories in the next chapter.

To conclude this section, information processes in HECA are similar to the basic units in Dennett's Pandemonium architecture—they are autonomous agents engaged in response-selection. However, HECA develops important general features concerning these autonomous information processes. Specifically, according to the Hierarchical Structure Thesis, hierarchical neural mechanisms represent the causal structures of the world at different depths, and/or process information with computations of different complexities. As a result, they subserve different levels of information processes that exhibit the tradeoff between flexibility vs. speed, on the one hand, and flexibility vs. capacity on the other. That is, higher-level information processes are slower and more flexible, and there are fewer numbers of them at a given time; conversely, lower-level information processes are faster and less flexible, and there are more numbers of them at any instance. Speed, capacity, and flexibility are all important considerations in the operation of dynamical cognitive systems. As a result, the Hierarchical Structure Thesis will have important implications for the control problem. I will discuss this issue in detail in the last section in the following chapter, as well as in Chapters 6 and 7.

3.3. The Cooperative Decision Thesis

In this section, I focus on the third thesis of HECA, cooperative and dynamical response-selection. Similar to Dennett's Pandemonium architecture, HECA utilizes a simple-message strategy for coordination; however, it revises Dennett's emphasis on competition and power struggle among demons, and conceives of response-selection from the viewpoint of epistemic cooperation. In the following, I will first state and explain the Cooperative Decision Thesis. Then, I will discuss the computational theory for this response-selection task, the *sequential probability ratio test* (SPRT). I will review the supporting empirical theory in the next chapter.

The thesis can be stated concisely:

> **Cooperative Decision Thesis**: intelligent decisions emerge from dynamical and epistemically cooperative processes of accumulating DVs from heterogeneous information processes until a decision threshold.

The decision process is described as follows: At various loci of the cognitive system, action-specifying information processes of various levels constantly specify conflicting response options. At the same time, action-evaluating information processes also evaluate the optimality of these options for the subject, and produce DVs to represent their evaluations. Each action-evaluating information process, depending on its algorithms and the availability and saliency of the relevant information, may evaluate different subsets of the total available information in different ways. Because information processes may start at different times and have different speeds, they will produce DVs at different times. The DVs produced for each option (in the form of positive or negative neural activation) are respectively integrated/accumulated in the associated neural population in time to reflect the collective evaluation for each response option. The DVs will continue to accumulate until one response option obtains enough DVs to reach a decision threshold. At this point, the response option is selected for execution (as illustrated in Figure 4.2). The autonomous information processes can be seen as cooperating epistemically and dynamically with each other to select the best response option for the subject.[72]

[72] For a recent literature that also draws the analogy between voting and neural information processing implementing SPRT, see (Marshall, 2011).

Moreover, cooperative response-selection is a general feature of the cognitive system; that is, all responses are selected this way. Not only are specific motor actions, such as raising one's arm, the results of cooperative selection processes, so too are more abstract goals that are not directly related to any specific motor movements, e.g., choosing to eat a banana rather than an apple. Also, cognitive actions at different levels of various hierarchies are selected this way, including the retrieval of information in long-term memory, updating of one's working memory, and selective routing of visual information.[73]

Let me clarify some of the key terms. First, response-selection in HECA is a dynamical, rather than static process, because time is one of the important parameters or variables in describing the processes. The DVs are generated and accumulated at different temporal points and throughout the deliberation time. Moreover, the exact time a certain piece of evidence is acquired and evaluated makes a difference to the final decision made. For example, suppose that all the positive DVs for the option B (Figure 4.2) arrives early, between 0 to 150ms, and all the negative evidence arrives later after 150ms. Without changing the total amount of positive and negative evidence, it may result in option B, instead of option A, crossing the threshold first and being selected for execution. That is, changing the dynamics of DVs can change the final decision made.

Second, response-selection in HECA is an epistemically cooperative, rather than competitive, process. Dennett's political metaphor of hegemony formation often invites an interpretation that decision-making is more about *power struggles* between coalitions of demons who pursue their own individual demon-interests than about the optimality of the decision for the subject. In HECA, in contrast, the response-selection is based on *evidence of the optimality* of the response options for the subject. Specifically, DVs are generated by action-evaluating information processes when they evaluate the options for their expected benefits for the subject (e.g., expected subjective utilities). Although different information processes may evaluate the response options differently (e.g., against different goals/standards, and consider different subsets of the total relevant information, etc.), they are different ways of pursuing the same goal (tracking the best option for the subject). Because they pursue a common goal through some sort of division of labor, the response-selection process in HECA can be seen as an epistemic cooperation.

Finally, intelligent decision-making (or specifically, the competence to choose the optimal response option reliably across different contexts) is an emergent property of the cooperative response-selection process, because the competence is a high-level (in the part-whole sense) capacity that is

---

[73] The selection processes described here, although prevalent in human cognitive systems, are unlikely to correspond directly to the deliberation processes we are most familiar with at the personal level, such as conscious reasoning and decision-making. As Susan Hurley argues (2001, p. 8), we need to avoid confusing claims about mental processes at the personal level (as those in our folk psychological theories) with claims about mental processes at the sub-personal level (such as in the cognitive architecture of scientific theories). In particular, we need to be wary of a naïve assumption of isomorphism between processes at the personal and sub-personal levels. For example, the common personal-level phenomenon of conscious deliberation about goals of, say, becoming a competent philosopher of cognitive science, are unlikely to be mapped isomorphically onto the sub-personal selection processes in HECA. For the purpose of this chapter, I am leaving open how the complex relation between them may be realized exactly. However, I speculate that it may be similar to Dennett's proposal that conscious thinking is a kind of skillful self-stimulation, which involves several cycles of internalized speech-acts (Dennett, 1991). That is, each internalized speech-act is the result of various selection processes (including ones involved in speech production) across the cognitive system, and it will prompt and nudge another wave of selection processes, which lead to the next internalized speech-act. In sum, a skillful (rational or reasonable) conscious deliberation process involves cycles of well-orchestrated selection process across a cognitive system. Also, see (Shadlen & Kiani, 2013, p. 800) for a similar account of the conscious will from the perspective of the dynamical decision-making processes.

the result of complex interactions among low-level component information processes, none of which possess the emergent competence.

Earlier I used the metaphor of a big family making collective decisions to elicit the intuition that epistemic cooperation can produce intelligent decisions. In what follows, I will briefly discuss the computational theory implemented by the cooperative response-selection process: the sequential probability ratio test (SPRT). Although SPRT is a very idealized rational model for decision-making, it will be adequate for my purpose of making more explicit how the response-selection processes enable emergent intelligence through dynamical epistemic cooperation.

Before I discuss the new dynamical vision of SPRT in more detail, I will have to briefly outline two more traditional and static rational models of human decision-making. The first is the expected utility theory (EUT) for the value-based decision, and the second is the signal detection theory (SDT), a Bayesian theory for perceptual decision-making.

According to one version of the EUT, the subjective EUT (Oppenheimer & Kelso, 2015), response options are selected based on their subjective expected utilities (SEU), such that:

$$\text{SEU} = \sum u(x_i)P(x_i)$$

$P(x_i)$ represents the subjective probability of one possible outcome (x_i) of a response option; and $u(x_i)$ represents the subjective utilities (subjective costs and benefits) that can be attributed to this possible outcome (x_i). The SEU of this response option is thus the sum of the product, i.e., $u(x_i)P(x_i)$, for each of its possible outcomes. The SEU of a response option, under the decision-making framework, is also called a decision variable (DV), which is a quantity transformed by a decision rule to produce a choice. The decision rule determines how the DVs are transformed into a commitment to a response option. For the simplest case, the decision rule is to choose the response option with the highest DV.

Note here the atemporal nature of the EUT as a computational theory of decision-making. Even though the information processing involved in calculating SEUs certainly takes time, the theory itself does not have any temporal parameters. Moreover, the most typical way of implementing EUT, the central systems of the Standard Account, does not allow an interpretation of intelligence as emergent from epistemic cooperation. This is because the relevant competence is directly implemented by a single system isomorphically, rather than achieved through the complex interaction of a group of information processes, none of which possess such competence individually.

Finally, EUT faces a large number of observed anomalies and is generally considered a failed paradigm for explaining human behavior (Oppenheimer & Kelso, 2015, pp. 280–2). These anomalies include loss aversion, preference reversal under time pressure, similarity effect, attraction effect, and compromise effect, etc. They share a commonality—that is, features of a decision task that are irrelevant to the assessment of SEUs of the best options affect the choice the subjects make (Busemeyer & Johnson, 2004, p. 139). For example, adding an inferior option (with less SEU than that of the best options) can make the subjects change their final decision from one of the best options to another one with the same (also highest) SEU. Although within the EUT framework, there are ways to accommodate these anomalies, e.g., modifying or adding parameters, positing complex subjective utility function, etc. Nevertheless, "[a]s complexity of the models continues to grow, so too does the number of observed anomalies that needs to be accounted for. This has led in the past half century to an expansion of decision models that can be considered

modifications of EUT" (Oppenheimer & Kelso, 2015, p. 282). Instead of making progress, EUT starts to look more and more like a research paradigm in crisis.[74]

Let's now turn to the second static account: signal detection theory (SDT). SDT is a Bayesian theory for perceptual decision-making. SDT provides a procedure to convert a single observation of evidence into a response in the context of a perceptual decision task. In a perceptual decision task, the goal has been fixed to select the response that indicates the current state of the world (or one of the possible type of stimuli presented). I need to introduce SDT in addition to EUT here because the sequential probability ratio test (SPRT) can be more easily understood if we see it as a dynamical extension of SDT.

To illustrate the SDT model, let me use the random dot motion discrimination task as an example: In this task, the subject is presented with a visual stimulus of a population of dots moving mostly randomly, but with a small trend of dots moving in one of the two possible opposing directions. For the purposes of the task, the subject needs to identify which one of the two possible directions the dots are moving towards.[75] Because there are only two possible directions of dot movement, we can think of the decision as involving a choice of hypotheses h_1 (the dots are moving left) and h_2 (the dots are moving right). The decision is based on the probability of h_1 and h_2, given the evidence (e) observed through visual perception (i.e., the posterior probability). That is, one should choose the hypothesis (and the relevant response) with the higher posterior probability (Gold & Shadlen, 2007). Based on Bayes' rule:

$$P(h_i|e) = P(h_i) \times P(e|h_i)/P(e)$$

$P(h_i|e)$ is the posteriors— the probability of a hypothesis (h_i) being true, given a piece of evidence (e) is observed. $P(h_i)$ is the prior, referring to the probability of h_i being true before obtaining the current evidence (e), and can been seen as a prediction based on information the agent possesses about the world prior to obtaining the evidence (e). $P(e|h_i)$ is the likelihood, the probability of observing the evidence (e), given a hypothesis (h_i) is true. The likelihood indicates the compatibility of the evidence (e) with a hypothesis (h_i). Here, we will ignore $P(e)$, the marginal likelihood, as it is the same for both hypotheses and as a result, does not influence the selection of them.

The decision variables (DVs) and decision rule for this simple two-alternative perceptual decision task can be represented as the following form (based on the ratio of the logarithms of the posteriors for the two hypotheses):[76]

---

[74] A prominent paradigm, heuristic decision-making (Gigerenzer & Gaissmaier, 2011) aims to address these anomalies with simple, domain-specific, decision-making algorithms, i.e., the heuristics. However, this approach faces a similar crisis: when increasing numbers of domain-specific heuristics are postulated to explain the anomalies, they start to seem *post hoc* and limited in their predictive power (Oppenheimer & Kelso, 2015, p. 282).

[75] When applying SDT to the analysis of the random dot motion paradigm, we take the temporally-extended presentation of stimuli as a single presentation of evidence, and the subject is only allowed to make a decision at the end of the presentation. This way, the significant temporal dimension of this task is compromised.

[76] The alternative representation of decision variables and decision rule is introduced for the convenience of explaining SPRT in the following paragraph.

$$DV = \log\frac{P(h_1|e)}{P(h_2|e)} = \log\frac{P(h_1)\times P(e|h_1)}{P(h_2)\times P(e|h_2)}$$

$$\text{Decision Rule: } \begin{cases} \text{If DV}>0\text{, choose } h_1 \\ \text{If DV}=0\text{, no decision} \\ \text{If DV}<0\text{, choose } h_2 \end{cases}$$

As we can see, SDT provides a Bayesian framework to form decisions that take into account a variety of information, including priors and evidence. However, it is nonetheless a static vision of human decision-making, because no temporal parameters are present in the framework. Similarly, this procedure, as implemented by the central systems, is also not readily interpreted as either an epistemic cooperation or an emergent phenomenon.

We are now in a position to consider the sequential probability ratio test (SPRT), a new account of human decision-making. SPRT can be viewed as a dynamical extension of the SDT, and it can accommodate a series of evidence from different sources acquired over time. Similarly, SPRT conceptualizes the decision process into the construction of DVs and the selection of a response based on the decision rule. The construction of DV is the same as that of the SDT framework. However, the decision rule differs in order to allow for a sequence of evidence acquisitions and evaluations. At each step, the decision rule determines whether it is time to stop the deliberation process and commit to a particular hypothesis, or whether no commitment can be made and more evidence is needed. The SPRT framework is introduced here in discrete sequences. However, we can imagine the time course of each sequence approaches the limit, and we get a continuous SPRT.

The process of SPRT can be represented schematically as Figure 4.4. The decision is based on a sequence of evidence acquisition. After each step of acquisition, a function, $f(e)$, converts the evidence acquired up until that point into a DV, and the decision rule will determine whether to commit to a particular hypothesis or to acquire more evidence.

$$\begin{array}{lllll} e_0 \rightarrow f_0(e_0) & \Rightarrow \underset{or}{Stop} & & & \\ & \swarrow & & & \\ e_1 \rightarrow & f_1(e_0,e_1) & \Rightarrow \underset{or}{Stop} & & \\ & & \swarrow & & \\ & & e_2 \rightarrow & f_1(e_0,e_1,e_2) & \Rightarrow \underset{or}{Stop} \\ & & & & \vdots \\ & & & e_n \rightarrow & f_1(e_0,e_1,\ldots,e_n) \Rightarrow \ldots \end{array}$$

Figure 4.4 The general framework of SPRT. e_0 can be interpreted as the prior probability's bearing on the hypothesis testing. Reprinted from (Gold & Shadlen, 2007).

Specifically, the decision variable constructed from multiple, independent pieces of evidence $(e_1, e_2, \ldots, e_n)$ can be represented (per SDT) as such:[77]

$$DV = \log\frac{P(h_1)\times P(e_1, e_2, \ldots, e_n|h_1)}{P(h_2)\times P(e_1, e_2, \ldots, e_n|h_2)}$$

[77] Evidence E and Evidence F are independent, both with respect to a state of affair X and with respect to its negation ¬X iff P(E∧F|X) = P(E|X) P(F|X) and P(E∧F|¬X) = P(E|¬X) P(F|¬X).

This construction of DV as the logarithm of the posterior probability ratio is equivalent to the following construction of DV as the cumulative sum of the logarithm of the prior probability ratio and that of the likelihood ratio of each piece of evidence:

$$\mathrm{DV} = \log\frac{\mathrm{P}(\mathrm{h}_1)}{\mathrm{P}(\mathrm{h}_2)} + \sum_{\mathrm{i}=1}^{\mathrm{n}} \log\frac{\mathrm{P}(\mathrm{e}_\mathrm{i}|\mathrm{h}_1)}{\mathrm{P}(\mathrm{e}_\mathrm{i}|\mathrm{h}_2)}$$

The decision rule here is to update the DV with new evidence until it reaches a positive criterion A (choose h_1) or negative criterion – A (choose h_2):

$$\textit{Decision Rule}: \begin{cases} \textit{If } DV \geq A, \textit{ choose } h_1 \\ \textit{If } -A < DV < A, \textit{ continue collecting evidence} \\ \textit{If } DV \leq -A, \textit{ choose } h_2 \end{cases}$$

The desired reliability is set by the positive and negative criteria of the decision rule, and can be controlled dynamically. The further away they are from zero, the higher the reliability of the final decision (because the posterior for the chosen hypothesis will be higher). Moreover, because collecting and evaluating more evidence will take more time and resources, the decision threshold also determines both the deliberation time taken before committing to a response option and the cognitive resources devoted to this decision. For example, a higher decision threshold is likely to lead to a more reliable, but slower and more expensive decision; conversely, a lower decision threshold will lead to a less reliable but faster and cheap decision. In other words, the decision threshold controls the "speed-accuracy" and "resource-accuracy" trade-off. As a result, a cognitive system can regulate the threshold to achieve the right balance between selecting reliable responses and meeting time and resource pressures. That is, SPRT addresses the important trade-offs that are crucial for animals solving problems in ecologically valid environments.[78]

This dynamical process of SPRT can be represented as the Figure 4.5.

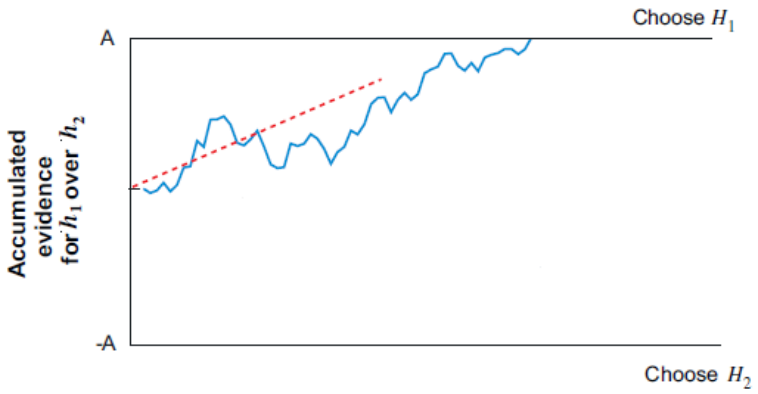

Figure 4.5 The dynamical process of SPRT. The horizontal axis represents the deliberation time. The vertical axis represents the amount of DV accumulated. The upper and lower bounds represent the positive and negative criteria in the decision rule. The deliberation process will continue to update the DV with new evidence until it reaches one of the bound, at which time a decision is made. Adapted from (Gold & Shadlen, 2007).

[78] SPRT has been proven to optimize the speed for decision of a required reliability (Bogacz, 2007, p. 119). That is, SPRT takes the shortest time to make a decision of a desired reliability, on average.

Now, we can get a rough sense of how the response-selection process in HECA is a dynamical and epistemically cooperative process from which intelligence emerges:

It is a dynamical process because it implements SPRT, which represents a truly dynamical vision of decision-making. Unlike the EUT and SDT frameworks described above, there is an explicit temporal dimension in SPRT. Time is an important parameter in this framework: The evidence is acquired and evaluated at different temporal points and continuously throughout the deliberation time. Moreover, the exact time when a certain piece of evidence is acquired and evaluated makes a difference to the final decision made.

More importantly, by showing that response-selection in HECA implements SPRT, it highlights not only the selection process's dynamical nature, but also its rationality—in particular, it emphasizes how epistemic cooperation enables intelligent decisions to emerge. Specifically, the DVs in SPRT ($\log\frac{P(h_1)}{P(h_2)}$ and $\log\frac{P(e_i|h_1)}{P(e_i|h_2)}$) can be calculated by different information processes. That is, different information processes evaluate different pieces of evidence in different ways to generate DVs. Accumulating the DVs from different information processes until a set threshold can increase the overall reliability of the final decision: because the decision threshold determines the amount of integrated DVs needed before committing to a particular response option (i.e., $\log\frac{P(h_1|e)}{P(h_2|e)}$), it establishes the level of expected reliability required before committing to a response option. The further away from zero the threshold is, the higher the expected reliability will be. In short, SPRT illustrates how the competence of reliable response-selection emerges from the epistemic cooperation of information processes of more limited computational power.

Before we conclude this section, there are some qualifications worth noting: First, SPRT applies to only perceptual decisions of two alternative choices. However, most of our more realistic decision-making involves (1) deliberations of more than two alternatives, and (2) value-based deliberation (i.e., there is a utility function or cost function to consider). Second, the version of SPRT I introduced above has an assumption of independence of evidence that realistic decision-making may not observe. Third, SPRT does not explicitly consider the heterogeneity of information processes that compute evidence to generate DVs and, specifically, the difference in the reliability of those information processes and the resulting difference in accuracy of the DVs they produce. The accuracy of DVs depends on the computational sophistication of the action-evaluating information processes and the amount of information they take into account. Finally, in the SPRT, both the construction of DVs and the decision rule are fixed and do not change over time. However, in the more general framework of sequential analysis, of which SPRT is a special case, both the construction of DVs and the decision rule can be adjusted and controlled over time. Human cognitive systems may adopt this more general framework in order to incorporate the cost of elapsed time—by lowering, for example, the criterion for decisions as time passes, and other considerations (Gold & Shadlen, 2007, p. 540).

There have been efforts to generalize SPRT for decisions involving more than two alternative choices, for example, research on the multi-hypothesis SPRT (McMillen & Holmes, 2006), and its implementation in the brain (Bogacz & Gurney, 2007). In addition, some cognitive scientists are confident that utility can be incorporated into SPRT: despite disagreement about how exactly it should be done, "value has been integrated into signal detection theory and all mathematical formalisms of decision theory in economics" (Shadlen & Kiani, 2013, p. 799). My hunch is that while much more work needs to be done, the core dynamical vision of decision-making embodied by SPRT is likely to remain intact.

I have shown that there is an alternative rational model available at the computational level of analysis of decision-making other than the good old-fashioned EUT and SDT implemented by the more traditional cognitive architecture (Fodor, 1983). Although the implementation of SPRT in human cognitive systems is inevitably approximate,[79] it explicitly demonstrates the epistemic foundation of the dynamical and cooperative response-selection.

To conclude this section, the Cooperative Decision Thesis develops the simple message-passing strategy of the embodied cognitive science approach, and revises the hegemony formation in Dennett's Pandemonium architecture by focusing on the cooperative aspect of response-selection. I have highlighted the competition's rationality, dynamical nature, and its generality in the cognitive system. I will discuss the empirical support and mechanistic detail for this thesis in the next chapter.

4. Conclusion: the (Left Leaning) Middle Way

In this chapter, I have formulated HECA, an empirical update of the Society of Mind account. I have articulated its conceptual shape by developing its three associated theses: the Embodied Agent Thesis, the Hierarchical Structure Thesis, and the Cooperative Decision Thesis. According to HECA, embodied information processes, subserved by hierarchical sensory and motor mechanisms, are the basic agents of the mind. In particular, the lower-level information processes have some autonomy from the higher-level processes because the former are not dictated by or under complete control of the latter in what they can or cannot do. Instead, action-specifying and action-evaluating information processes of different levels cooperate during response-selection to determine which internal or external responses will be executed. Response options are specified continuously by action-specifying information processes. Additionally, action-evaluating information processes in favor of a given option continuously provide positive DVs (in the form of neural activation); in contrast, action-evaluating information processes against this given option provide negative DVs (in the form of neural inhibition). The DVs are accumulated in the associated neural populations to reflect the collective evaluation of response options. When the accumulated activation for an option passes the threshold for decision, the option is selected for execution.

What remains to be seen is how HECA compares to existing models of cognitive architecture. In this section, I will show that, roughly, if we place these theories on a spectrum ranging from classical to anti-classical theories, the hierarchical competition model is situated somewhere in the middle, though farther toward the anti-classical side. In Section 4.1, I will begin by contrasting HECA with the classical cognitive science's Standard Account to show how it rejects all three assumptions of this "right-wing" theory of human cognition. In Section 4.2, I will locate the model on the "center-left" of the cognitive architecture spectrum. I shall do so by establishing HECA's compatibility with the moderate version of the embodied mind thesis, as well as its incompatibility with the thesis of radical embodiment.

4.1. Anti-classical Cognitive Architecture

Similar to Dennett's pandemonium account, HECA is an anti-classical model, as it rejects all three features of the classical cognitive architecture discussed in Chapter 1. First, HECA's information processing mechanisms do not separate into peripheral perception and motor modules, with central modules mediating between them. Instead, embodied information processes subserved by both

[79] The implementation of SPRT in human cognitive systems has to be approximate, at least most of the time, because the computation required for SPRT, similar to that required for EUT, become intractable very quickly when the information involved increases. I discussed the relevant issue in Chapter 2.

perceptual and motor mechanisms form the basic units that function without the mediation of central cognitive systems.

Second, HECA has no central systems that mediate between or dictate to the dumb perceptual and motor modules. Instead of a single source of intelligence, there are only different levels of autonomous embodied information processes, which interact with each other in the production of intelligent behavior.

Finally, the capacities related to central cognition in HECA are not implemented by classical computational processes. Rather, the empirical theories we will review in the next chapter suggest that the information processes and competences underlying central cognition (and cognition in the broader sense) are implemented in action-centric, modality-specific representations and processes. Even if the higher-level mechanisms tend to involve more abstract representations, they will not involve the kind of amodal and action-neutral representations required for the classical computational architecture.

One implication follows from the rejection of the classical architecture. HECA allows for a flexible trade-off between producing fast, inexpensive, but likely less optimal responses on the one hand, and slow, expensive, but likely more optimal responses on the other. This trade-off is crucial for meeting real-time ecological constraints. The classical cognitive architecture lacks this valuable feature due to what Rodney Brooks calls the problem of "representational bottleneck." This problem is generated by positing a central executive planner that is "privy to all the information available anywhere in the system"(Clark, 1998, p. 21). As Clark explains:

> The reason [for this problem] is that the incoming sensory information must be converted into a single symbolic code so that such a planner can deal with it. And the planners' output will itself have to be converted from its propriety code into the various formats needed to control various types of motor response. These steps of translation are time-consuming and expensive. (Clark, 1998, p. 21)

Instead, response-selection in HECA can involve different levels of information processes flexibly. If time and resources are constrained, response-selection can be restricted to fast and large-capacity lower-level processes; if the accuracy and optimality of responses are more important considerations, more flexible higher-level processes can be recruited to participate in the response-selection. This raises the question of how the tradeoff is intelligently managed, which I will discuss in the next chapter.

4.2. Moderate Embodiment

HECA is situated on the left-leaning center of the spectrum of cognitive architecture. This is the case because, despite being a brain-centric model, HECA supports, or at minimum is compatible with, many moderate versions of the 4E (embodied, enactive, embedded, and extended) cognition. It is not, however, compatible with some of the most radical versions of 4E theories.

My model directly supports a moderate version of embodied cognition, according to which cognitive processes (in both the narrow and general senses) constitutively depend on the sensorimotor parts of the brain. This implication can be drawn directly from the manner in which central cognitive capacities emerge out of sensorimotor processes at different levels in my model.

In addition, HECA is also compatible with, and indirectly supports, stronger versions of embodied cognition, according to which cognition depends, causally and/or constitutively, on the (non-

neural) body and even the environment. The dynamical nature of the sensorimotor information processes lends support to the coupling between the brain and body, and perhaps the brain, body, and environment, a requisite for the constitutive dependence of cognition on the body and the environment.

Finally, my model is not compatible with radical embodied cognition, according to which representations of any sorts play little to no role in cognition. In fact, the hierarchical competition model is a "representation-hungry" model: despite the rejection of classical computationalism and rich internal models of the world, its hierarchical mechanisms do utilize modality-specific and action-centric representations, as well as multiple, distributed, partial, and task-specific internal models of the world.[80]

In this chapter, I've laid out the conceptual shape of HECA and situated it on the center-left of the spectrum of cognitive architectures. In the next chapter, I will discuss the current empirical theories supporting HECA and the remaining issues challenging it. As we will see, the empirical advance may help create as many problems as it solves.

[80] In this regard, my model is similar to, and compatible with, the massively representational view of mind proposed by Robert Rupert (2011).

5

Society of Mind in the Twenty-First Century II: Empirical Support, Progress, and Remaining Challenges

1. Introduction

In the previous chapter, I laid out the conceptual shape of the hierarchical embodied cooperative architecture (HECA). HECA updates Dennett's Pandemonium architecture by incorporating insights from empirical literature developed in recent years. Similar to Dennett's account, HECA posits autonomous action-specifying and action-evaluating information processes that interact with each other to select actions. In developing HECA, I have advanced three new theses. First, the Embodied Agent Thesis, according to which autonomous action-specifying and action-evaluating information processes are embodied. As we saw, this thesis involves three claims about information processes: (1) the universal constitutive dependence on offline embodiment, (2) the existential constitutive dependence on online embodiment, as well as (3) the existential representational embodiment. Second, the Hierarchical Structure Thesis states that neural mechanisms are structured hierarchically, and they represent the causal structures of the world (including the agent) at different depths and/or process information with computations of different complexities. As a result, these mechanisms subserve different levels of information processes that vary systematically in their flexibility, capacity, and speed. Last, the Cooperative Decision Thesis asserts that intelligent decisions emerge from dynamical and epistemically cooperative processes of accumulating decision variables (DVs) from heterogeneous information processes until a decision threshold.

I've claimed that recent large-scope empirical theories of cognition support these theses. In this chapter, I shall substantiate my claim by reviewing these theories and demonstrating relevant supporting evidence. Furthermore, the relevant computational and mechanistic details of these models will help flesh out the core features introduced conceptually in the previous chapter.

Finally, they will illuminate how HECA addresses the control problem, as well as highlight new and as yet unresolved challenges.

In Section 2, I will first review the *affordance competition hypothesis* for embodied decision-making proposed by Paul Cisek and his colleagues (Cisek, 2012a, 2012b; Cisek & Kalaska, 2010; Cisek & Pastor-Bernier, 2014). Then, I will discuss *hierarchical models of perception and action-control* (in section 3) and the *model-based/mode-free reinforcement learning and control* literature (in section 4). Afterward, I will briefly discuss another two recent accounts of cognition, the *predictive mind* (in section 5) and *dual-process theories* (in section 6). Despite the fact that they are supportive of HECA, I've given them a secondary role; neither of them is, at present, mechanistically well-specified. Section 7 will discuss the *sequential sampling models* prominent in areas of decision neuroscience such as neuroeconomics. They provide empirical support specifically for the Cooperative Decision Thesis. In Section 8, I will discuss the progress HECA makes in solving the control problem, as well as the new challenges it poses for intelligently coordinating neural mechanisms.

2. The Affordance Competition Hypothesis

In this section, I will focus on one recent development, Cisek's *affordance competition hypothesis* (ACH) (Cisek, 2012a, 2012b; Cisek & Kalaska, 2010; Cisek & Pastor-Bernier, 2014). Cisek develops the ACH primarily as a model for what he calls "embodied decisions," which are "decisions between [motor] actions during ongoing activity"(Cisek & Pastor-Bernier, 2014, p. 3). This type of decision presents a challenge to cognitive systems because it requires the constant production of adaptive behaviors in response to continuously changing internal and external environments. Cisek believes his model can meet the challenge because it posits a functional architecture that continuously processes sensory information to specify potential actions currently available in the environment (the affordances) and, in parallel, select between them based on biasing information (the competition). Additionally, his model is supported by several lines of empirical evidence that are hard to reconcile with more traditional models that strictly distinguish between mechanisms of perception, cognition, and action (Cisek, 2012b, pp. 931–2). In the following, I briefly review how ACH supports HECA's three main theses and discuss its relevant computational and mechanistic details. I will trace the shared themes developed in these two models—embodied agents, cooperative decision, and hierarchical structure—but also discuss their distinctive features. I end by discussing one specific line of the empirical evidence for ACH.

2.1. Embodied Agent

Cisek's ACH supports the Embodied Agent Thesis. The agents in Cisek's model are functional mechanisms categorized into two main types: those involved in action-specification, and those involved in action-selection.[81] Similar to HECA's action-specifying information processes, the action-specification mechanisms (the sensorimotor loops within the frontoparietal system) process visual information and specify potential actions of various kinds (illustrated as the filled black arrows from the primary visual cortex to the premotor cortex in Figure 5.1).[82] Specifically, several parietal

[81] Instead of the traditional functional distinction of perceptual, cognitive, and motor mechanisms, Cisek believes the distinction between action-specification and action-selection mechanisms is more suitable for understanding cognitive systems, except for the purpose of understanding primary sensory and motor regions, where the concepts of purely perceptual and motor mechanisms may still apply fruitfully (Cisek, 2012a, p. 228).

[82] The examples used will draw heavily on visually-guided actions; however, the model should apply to motor actions across the board.

cortex areas within the "dorsal stream" of the visual system (Goodale & Milner, 1992), which is traditionally thought to specialize in processing visual information for guiding various types of actions, are strongly and reciprocally connected with corresponding frontal cortex areas considered to specialize in complementary motor functions. The frontoparietal system constitutes a set of sensorimotor loops, each of which specifies different aspects of an action, such as potential eye saccade movement and possible directions for reaching.

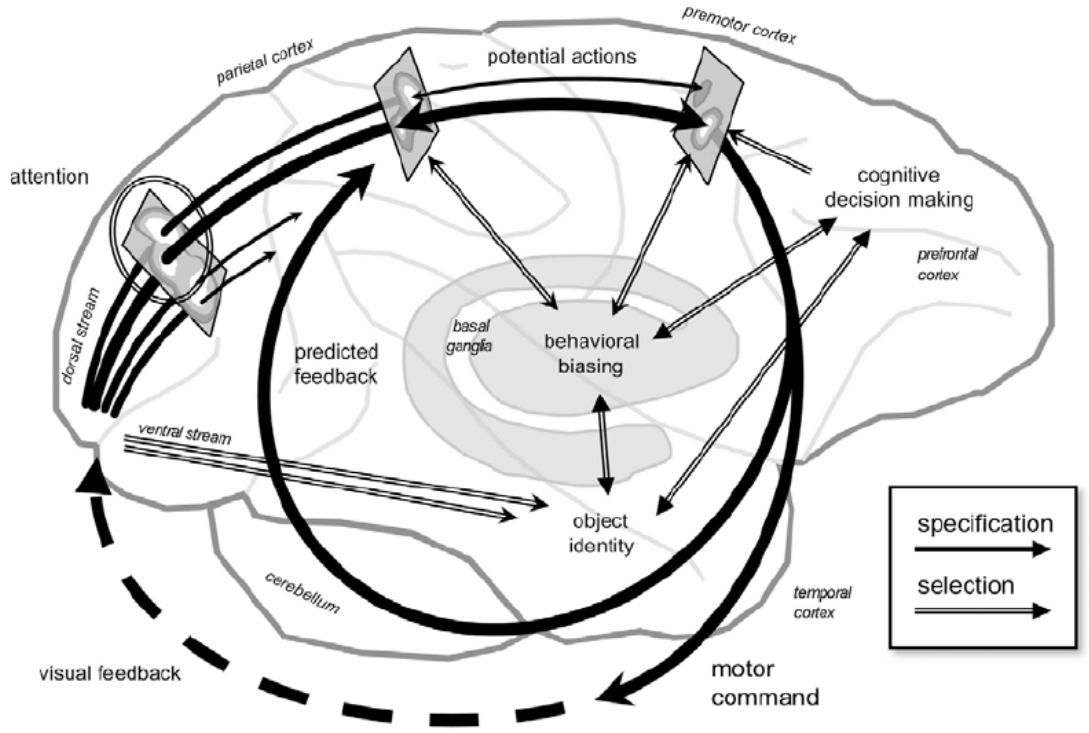

Figure 5.1 Sketch of major components of the *ACH* in a visually-guided action production. Filled black arrows represent processes of action specification (sensorimotor loops) in the dorsal stream, transforming sensory information in the visual cortex into representations of potential actions. The filled dark arrow through cerebellum and the dashed black arrow through the environment reflect further and more fine-grained action-specification through internal and external feedbacks. The double-line arrows represent the biasing influences of action-selection mechanisms from various brain regions. Figure excerpted from (Cisek, 2012a).

Together, these loops form a "sensorimotor map" that represents a set of potential actions made available by the environment (affordances according to Cisek[83]) to select from.[84] This set of potential actions (and their parameters) are represented as patterns of activity in tuned neural populations. However, the neural populations do not just represent the value of a particular

[83] Note that it is at least imprecise to call competing sensorimotor representations of potential actions "affordance," because the proper definition of "affordance" is "the perceivable relationship between an organism's abilities and features of the environment" (Anderson, 2014, p. 218).

[84] Note that not all currently available actions will be represented in the dorsal stream; some of the available actions may fail to be registered by the cognitive system and some may be eliminated early in the information processing. Also, more fine-grained action-specification seems to involve other mechanisms outside of the dorsal stream, e.g., one that involves overt feedback from the environment (the dashed black arrow in Figure 5.1) and one that involves internal feedback from cerebellum (the filled dark arrow running across cerebellum in Figure 5.1) (Cisek, 2012a, p. 211).

parameter of *one single* potential action (e.g., a reaching movement's direction) but an entire distribution of that parameter's values of *all* potential actions (e.g., many potential reaching directions represented together). For example, Figure 5.2 illustrates a simplified model of a neural population of cells with a different tuning direction for reaching movement. Together, their patterns of activation can represent one reaching direction (top) or multiple mutually exclusive ones (bottom). Roughly, the levels of activations within the neural populations associated with potential actions reflect their strength respectively, and determine which action will be selected for execution.[85]

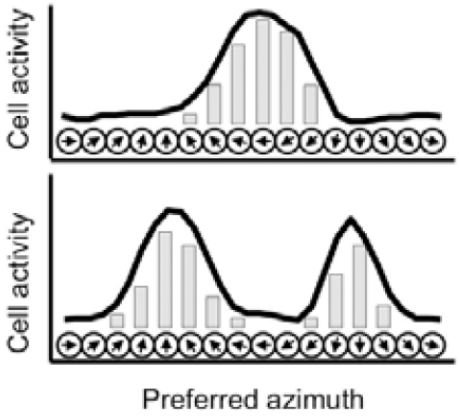

Figure 5.2 A neural population representing a single reaching direction (top) or multiple reaching directions (bottom). Figure excerpted from (Cisek, 2012a, p. 218).

Similar to HECA's action-evaluating information processes, ACH's action-selection mechanisms can also influence the levels of activation associated with potential actions. These mechanisms evaluate potential actions against various different criteria and provide biasing information to influence the outcome of competition between potential actions. For example, the ventral stream of the visual system responsible for object recognition (Goodale & Milner, 1992) works as an action-selection mechanism, evaluating the appropriateness of potential actions in the presence of particular objects. Based on the evaluation, it then increases or decreases the activation levels of neural populations associated with those potential actions respectively (Cisek, 2012a, p. 214). Figure 5.1 illustrates several action-selection mechanisms (represented by double-line arrows), exerting influence at different stages of the action-specification processes. Each of these action-selection mechanisms evaluate different aspects of the potential actions. The orbital frontal cortex, for example, evaluates the objective economic value of the outcome of a particular action, while the lateral prefrontal cortex evaluates the degree to which an action conforms to context-dependent rules operating currently (Cisek, 2012b, p. 930; Cisek & Pastor-Bernier, 2014, pp. 9–10).

In short, Cisek's ACH provides at least partial support for the Embodied Agent Thesis. The ACH suggests that the action-specifying agents (the sensorimotor loops) in the motor decision-making domains depend constitutively on sensorimotor mechanisms. As a result, ACH supports the "constitutive dependence on offline embodiment" claim in the motor domain: i.e., the capacity for motor decision-making depends constitutively on sensorimotor neural mechanisms. More interestingly, because ACH posits feedback loops operating via the body and the environment are

[85] Note that the functionally-individuated action-specification demons, each of which propose a particular action, are implemented in the same neural populations. It is a good example of how distinct entities at one level of analysis need not be implemented by distinct entities at another level of analysis.

involved in action-specification, it also suggests the "existential constitutive dependence on online embodiment" claim to be true: the motor decision-making capacity may depend constitutively on bodily or environmental mechanisms as well. However, the model does not make explicit claims about the embodied nature of the action-evaluating agents.

2.2. Cooperative Decision

Cisek's model also supports the Cooperative Decision Thesis. As we will see in the following, Cisek uses the concept of competition, instead of cooperation, to describe his model's response-selection process. However, it will soon be clear that on his model, the agents involved in the response-selection process have the common goal of selecting the best option for the subject, rather than the goal of fulfilling their respective interests. In addition, the decision mechanisms discussed in Cisek's model are similar to those employed in the sequential sampling models of decision-making (to be discussed in Section 7), which are mechanistic models that implement the sequential probability ratio test introduced in the last chapter. As a result, the response-selection in his model is best seen as implementing a cooperative process.

According to Cisek's ACH, the competition among agents occurs within the action-specification mechanisms, i.e., the sensorimotor map in the frontal-parietal area. Cells with different tuning preferences (representing parameters for different actions) within a neural population mutually inhibit each other, while cells with the same tuning preferences excite each other. The competition is further biased by action-selection mechanisms: information favoring a potential action excites the cells associated with an action, and information against the action inhibits the cells associated with it (Cisek, 2012a, pp. 213–4). In short, mechanisms in the same coalition excite each other, while mechanisms in competing coalitions inhibit each other. For example, the competition can be represented as the scheme (b) in Figure 5.3: Within the red box, two actions compete through mutual inhibition. Biasing information (e.g., action costs and outcome values) influence the competition by exciting or inhibiting the conflicting actions. An action is selected when a coalition of agents, the action-specification and action-selection mechanisms favoring the same action, become sufficiently strong and can conclusively suppress other competing coalitions' neural activities.[86] However, the mutual inhibition and excitation should be interpreted as an epistemic cooperation, rather than competition, because they implement the SPRT (as we will see in Section 7). In short, Cisek's ACH, despite its reference to competition, can be properly interpreted to support the Cooperative Decision Thesis.

2.3.Hierarchical Structure

Similar to the Pandemonium architecture, the ACH does not account for the selection of internal actions. This significantly limits its application because the selection of internal actions is an important issue for a complex cognitive system, such as the human cognitive system. Nevertheless, Cisek briefly addresses issues related to the selection of cognitive action in his less developed "distributed consensus model" (Cisek, 2012a, 2012b; Cisek & Pastor-Bernier, 2014). In this model, he explicitly generalizes the ACH to include the selection of higher-level goals at higher-level mechanisms. As a result, ACH's extension, the distributed consensus model, can be seen as support for the Hierarchical Structure Thesis.

---

[86] The competition process in Cisek's model is more complex than is reviewed here. For more detail, see the discussion of *se*quential sampling models of decision making, which will be discussed in Section 7.

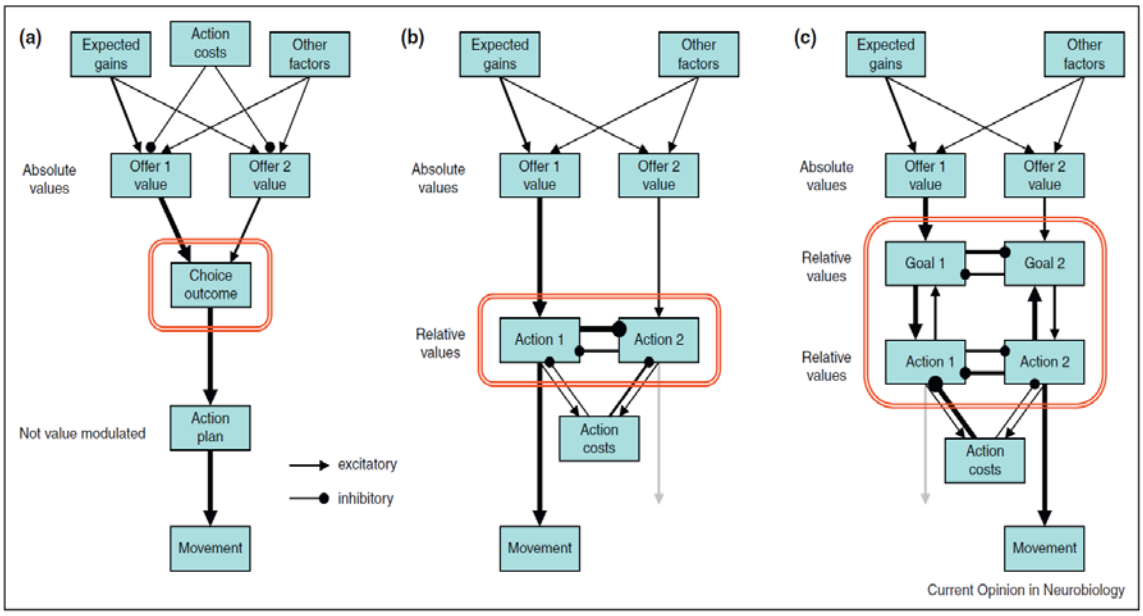

Figure 5.3 Three schemes of action-selection. The red box indicates the locus of selection and arrows represent excitation and inhibition with thickness representing strength. Scheme (a) represents the more traditional expected utility theory of action-selection; scheme (b) represents the *ACH* for motor action-selection; scheme (c) represents the *distributed consensus model* of action-selection. Reprinted from (Cisek, 2012b)

Briefly put, higher-level goals are those that do not correspond to specific movements (e.g., looking to the right or to the left), but have a bearing on the final movement selected. For example, choosing to obtain oranges rather than apples involves the selection of a higher-level goal that does not dictate a movement, but will ultimately influence the movement adopted to achieve this goal. In Figure 5.3, the scheme (c) illustrates how competition between higher-level goals interacts with the competition of lower-level action options in the production of motor actions: the competition at the higher-level can lead to the adoption of one goal over the other, which then biases the lower-level competition through top-down information. In contrast, the lower-level competition may settle first and influence higher-level goal selection through the bottom-up connection.

What Cisek implicitly suggests, but does not articulate or fully develop here, is a model where there are selections of (internal and external) actions at multiple levels in both the action-specification and action-selection mechanisms. In fact, the strict distinction between the action-selection and action-specification mechanisms disappear, because an action-selection mechanism for a motor action at the lower-level may be itself an action-specification mechanism for a higher-level cognitive action (e.g., the higher-order goal discussed above). What we have now is a model of mind with response-selection at various levels of hierarchies of information processes. The result of each response-selection will influence other competitions elsewhere; the final action reflects the result of cooperative epistemic activities distributed across the brain. Consequently, "the decision is not determined by any single central executive, but…a 'distributed consensus'" (Cisek & Pastor-Bernier, 2014, p. 4). To sum up, the *distributed consensus model*, an extension of the ACH, can be seen as support for the Hierarchical Structure Thesis.

Before we end this section, I would like to discuss just one line of empirical evidence for the ACH in order to illustrate some of the surprising empirical support it receives. Recall the bureaucratic model of the Standard Account discussed in Chapter 1, where decisions of "what to do" are made (1) in the central systems, and (2) *before* the decisions of "how to do" are made in the peripheral motor modules. Both ACH and HECA predict a different order. That is, the "what to do" decisions

are made (1) in the same brain area as, and (2) *after* or *at the same time as* the decisions of "how to do" are made (Cisek, 2012b, p. 211).

This is because the ACH posits that the specification of response options (the "how to do" decisions) starts earlier and continue as the evaluation and selection of response options (the "what to do" decisions) occurs in the same area, as Cisek puts it:

> The mechanisms that serve embodied decisions must process sensory information rapidly and continuously, specifying and re-specifying available actions in parallel while at the same time evaluating the options and deciding whether to persist in a given activity or switch to a new one. Thus, the temporal distinction between thinking about the choice and then implementing the response, so central to economic theory and laboratory experiments on decisions, simply does not apply to decisions made during interactive behaviour. (Cisek & Pastor-Bernier, 2014, p. 3)

In fact, empirical evidence in neuroeconomics supports this. For example, in an economic game involving primates selecting an option with an eye movement, the decision concerning "which superior option to choose" (the "what to do" decisions) are made at the same brain area of lateral intraparietal cortex (which is a motor area associated with controlling eye movement specifically) and at the same time as the decisions involving "which direction to move toward" (the "how to do" decisions) are made (Shadlen & Kiani, 2013, p. 797).[87]

In conclusion, the ACH supports HECA's all three theses. In the next four sections, we will review models that primarily support the Hierarchical Structure Thesis.

3. Hierarchical Models of Perception and Action-Control

The first set of models which strongly support the Hierarchical Structure Thesis are the *hierarchical models of perception and action-control* (HPA). Hierarchical models of perception have a longer history than those of action-control, and are rooted in physiological and neuropsychological research (Bond, 2004; Felleman & van Essen, 1991; van Essen, Felleman, DeYoe, Olavarria, & Knierim, 1990). However, recent fMRI and computational modeling studies have confirmed the existence of hierarchical structures in motor processing (Badre, 2008; Botvinick, Niv, & Barto, 2009; Uithol, van Rooij, Bekkering, & Haselager, 2012). Together, the empirical evidence suggests integrative perception-action hierarchies of human cognition (Bond, 2004; Fuster, 2004). In the following, I will review the common features of these models and their empirical support respectively.

Empirical researchers have shown that there are both information processing mechanisms structured hierarchically in our perceptual systems (Felleman & van Essen, 1991; Fuster, 2004) as well as action-control systems (Badre, 2008; Botvinick et al., 2009; Uithol et al., 2012). The visual system is one of the best understood mental mechanisms. Through physiological and neuropsychological research, we have acquired an extremely detailed understanding of the visual information-processing hierarchy (Bechtel, 2008, p. 89). For example, David van Essen and his colleagues have identified in the macaque's visual system thirty-two processing components arranged into a complex hierarchy (Felleman & van Essen, 1991; van Essen et al., 1990). (See Figure 5.4).

---

[87] For more details and the empirical inference and justification supporting this interpretation of the neuroscientific data, see (Cisek, 2012b; Cisek & Pastor-Bernier, 2014; Shadlen & Kiani, 2013, p. 797).

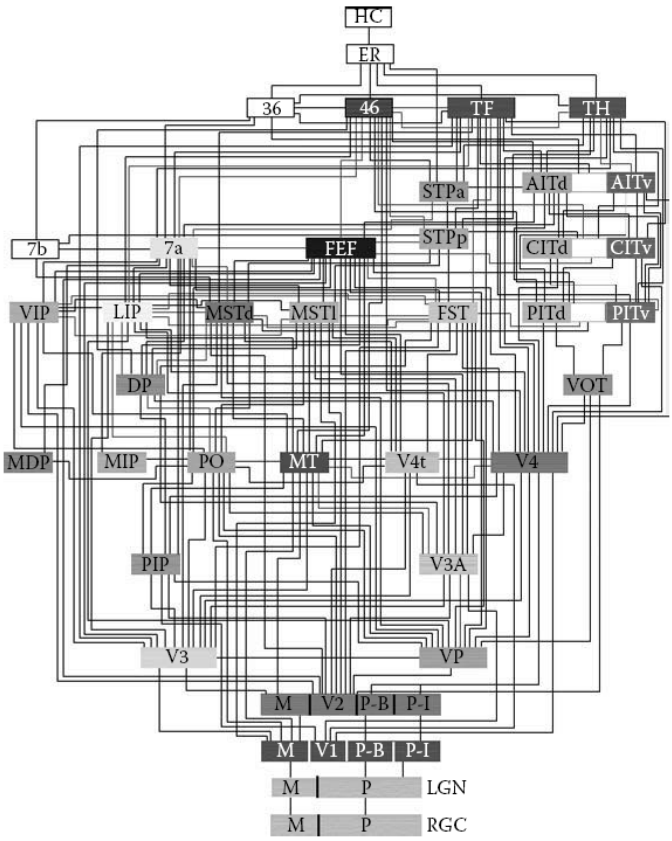

Figure 5.4 A representation of 32 cortical visual areas arranged into an information-processing hierarchy. Reprinted from Felleman & van Essen (1991).

Similarly, recent research also supports the existence of hierarchical structures in the action-control system (Botvinick, 2008; Koechlin, Ody, & Kouneiher, 2003; Picard & Strick, 1996; Uithol et al., 2012). For instance, Etienne Koechlin and his colleagues' fMRI study (2003) suggest the existence of hierarchical structures of action control (Figure 5.5) and localize them within the frontal cortex, from the premotor area (lower-level mechanisms) to the rostral lateral prefrontal area (higher-level ones).

The evidence for perceptual and action-control hierarchies, together with empirical findings for reciprocal connections between areas of the same rank within each hierarchy (Bond, 2004, p. 72) , have led some researchers to propose an integrative perception-action hierarchical model (Bond, 2004; Fuster, 2004) such as the one illustrated in Figure 4.1 (p.66).

This type of hierarchical model allows information processes to involve different levels of mechanisms in the production of internal and external actions. For example, a motor response can be generated by a lower-level information process that involves the primary sensory cortex, the unimodal association area, the premotor cortex, and the primary motor cortex. Alternatively, a motor response may also be a higher-level response generated by processes involving higher-level mechanisms in the polymodal association cortical area and rostral prefrontal cortex.

In addition, the HPA also supports the Hierarchical Structure Thesis' more specific claim about the systematic difference between higher-level and lower-level information processes with regard to their flexibility, capacity, and speed. In the following, I will review empirical support for these features, respectively.

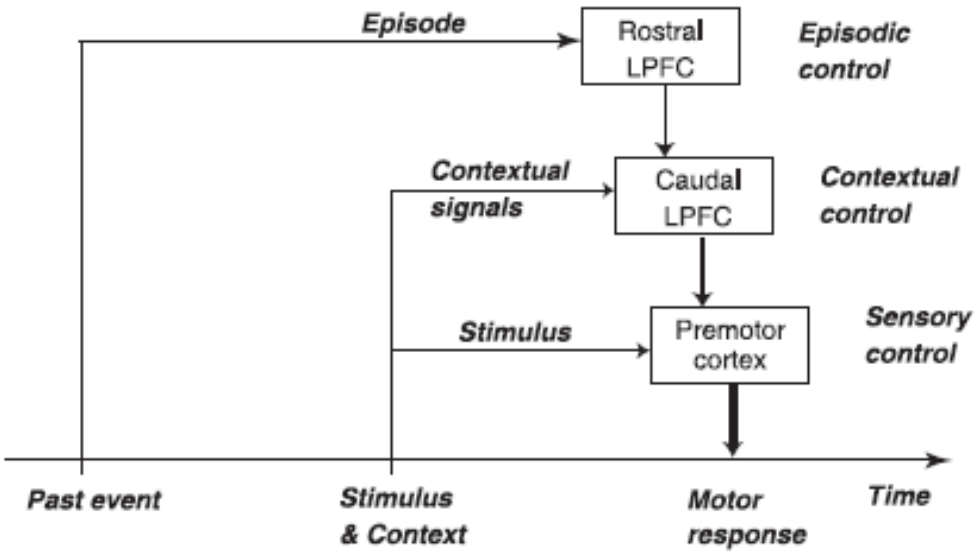

Figure 5.5 Three types of controls in three nested levels of information processing posited in Koechlin's functional model of cognitive control. Sensory control subserved by the premotor cortex receives specific sensory information and selects the appropriate motor action. The contextual control by the caudal lateral prefrontal cortex (LPFC) processes integrated contextual information of the environment and select the appropriate premotor representation (stimulus-response associations). Finally, the episodic control subserved by the rostral LPFC select the appropriate causal LPFC representations (the entire task sets of stimulus-response association) according to episodic information about events previously occurred. Reprinted from (Koechlin et al., 2003).

Flexibility

In HPA, information processes at different levels in the perception-action hierarchies also differ in their flexibility. Specifically, higher-level processes are more likely to specify more optimal responses and provide more accurate evidence in a wide range of contexts. This is because higher-level perceptual and motor mechanisms are required to process relevant information sensitive to the "complex and temporally remote stimuli" and "complex schemas or plans" in complex contexts (Fuster, 2004, p. 144) in order to generate optimal responses and accurate evidence.

In *HPA*, mechanisms from lower-levels to higher-levels deal with information processing of increasing temporal and spatial abstraction and integration (in perceptual hierarchies) or policy/goal abstraction and complexity (in motor hierarchies) (Botvinick, 2008; Fuster, 2004; Koechlin et al., 2003; Uithol et al., 2012). For example, as illustrated in Figure 5.5, Etienne Koechlin and his colleagues (2003, p. 1181) argue that three types of control are processed at three levels of hierarchical mechanisms. At the lowest level, the neural mechanisms in the premotor cortex are concerned with selecting specific and immediate motor actions—for example, determining whether to pick up and drink the beer when one sees a pitcher on the table. At the middle level, neural mechanisms in the caudal lateral prefrontal cortex (LPFC) deal with the action policy within one particular temporally and spatially extended context (i.e., the entire set of stimulus-response associations in the particular context). For example, they process information about whether it is appropriate to drink beer in a particular social situation. At the highest level, neural mechanisms in the rostral LPFC manage the changes between different contexts according to past events and present goals, e.g., determining what the current social context is, using episodic information about whether one is currently in one's own wedding banquet.

The difference in the abstraction of information processing among levels in action-control hierarchies is matched in the perception hierarchies. In lower-levels of perceptual hierarchies,

simple and concrete features of the world—including shapes, colors and concrete objects—are represented. As we ascend the hierarchy, we find that mechanisms begin to store more complex and abstract information about the world, such as situational social etiquette, as well as personal history with a particular place or individual.

We should note that the issue of how to best characterize the representations and information processing of the perception-action hierarchy remains a contentious one (O'Reilly, Herd, & Pauli, 2010; Uithol et al., 2012). However, this controversy will not matter for our purposes, because models agree that higher-level mechanisms deal with representations concerning the deeper structure of the world, and with information processing that is more complex. These models also agree that higher-level information processes are more flexible than lower-level ones.

Speed

With regard to the feature of speed, concrete empirical data shows that higher-level perceptual and motor processes tend to have slower reaction times, while lower-level processes tend to have faster reaction times. For example, Koechlin and his colleagues (2003) report significantly increased reaction times as information processing involves higher-level controls. Similarly, Jean Bullier, in a comprehensive review, notes a general trend of increased neuronal reaction time (associated with the registering of stimuli), from the early visual cortex, V1 (around 40 msec), to the orbitofrontal cortex (around 140 msec) (Bullier, 2001, p. 98; Gold & Shadlen, 2007, p. 558).[88]

Capacity

There is also concrete empirical data showing that there is a smaller capacity for higher-level perceptual and motor processes. The relative limitation in the numbers of higher-level information processes is best exemplified in the perceptual hierarchy. For instance, when we fixate the letter A in the center of Figure 5.6, all the surrounding letters are equally visible.[89] However, it is impossible to recognize all of the letters at the same time. In order to recognize the letters, we need to direct our visual attention to each individual letter or a small cluster of them at a time (van Essen et al., 1990, p. 17). Recognition tasks require relatively high-level mechanisms to process. Such mechanisms only have the capacity to process a selective portion of information received from lower-level mechanisms (Olshausen, Anderson, & Van Essen, 1993). This is why although we can "see" all the letters, we can only "recognize" a small subset of them. Similarly, such capacity-limitation can be found in the action-control hierarchy (Botvinick, 2008; Miller & Cohen, 2001). Matthew Botvinick (2008, p. 203) claims that working memory is required to activate and maintain representations in some higher-level processes. Because the working memory system is capacity-

---

[88] In this research, the briefest response time (from stimuli to neuronal responses of registering the stimuli) involving very high-level mechanisms, such as orbitofrontal cortex, is around 140 msec (Bullier, 2001, p. 98). This fact is significant because the response time for a simple perceptual-motor task, e.g., saccadic eye movement to a visual target, averages around 200 msec (Gold & Shadlen, 2007, p. 558). This suggests that, even for very fast motor responses, there is a window of time for higher-level mechanisms to be truly involved in the information processing, at least in some very cursory way.

Moreover, the variance of reaction time can be larger in more complex contexts. The above data reflects only relatively simple perceptual decisions such as reporting the sighting of a particular stimulus with the designated motor response.

[89] They are equally visible because the increasingly larger size of the letters in the outer circles compensates for the declining visual acuity toward the periphery of the visual field (van Essen, Felleman, DeYoe, Olavarria, & Knierim, 1990, p. 17) .

limited (Baddeley, 2007), it contributes to the capacity limitation of these higher-level processes. To sum up, HPA strongly supports HECA's Hierarchical Structure Thesis.[90]

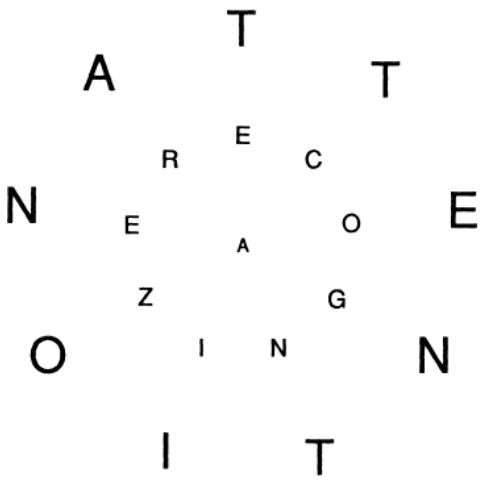

Figure 5.6 A visual display illustrates the capacity limitation of the higher-level visual processing. When looking at the center of the display, we can see every letter; however, we need to direct our visual attention to individual letters, or to a small cluster of letters, in order to recognize them. Reprinted from (van Essen et al., 1990).

4. Model-Based and Model-Free Reinforcement Learning and Control

A second major line of empirical literature that strongly supports the Hierarchical Structure Thesis is model-based and model-free reinforcement learning and control (MBMF). Although model-based and model-free reinforcement learning algorithms were initially developed from the normative account of instrumental control of behavior in the field of machine learning (Richard S. Sutton & Barto, 1998),[91] empirical neuroscientists have found increasing support for their neural

---

[90] I want to note that the autonomy of information processes is also supported by HPA, although it is often not explicitly expressed. For example, we often find claims contrasting "automatic" responses driven solely by lower-level processes, and "complex" responses that require the involvement of higher-level mechanisms:

> Automatic and well-rehearsed actions in response to simple stimuli are integrated at low levels of the cycle, in sensory areas of the posterior (perceptual) hierarchy and in motor areas of the frontal (executive) hierarchy. More complex behavior, guided by more complex and temporally remote stimuli, requires integration at higher cortical levels of both perceptual and executive hierarchies, namely areas of higher sensory association and prefrontal cortex. (Fuster, 2004, p. 144)

That is, lower-level information processes can autonomously drive responses without the involvement and endorsement of higher-level processes.

Moreover, according to the guided activation theory (Miller & Cohen, 2001), higher-level mechanisms (say, in the prefrontal cortex) provide top-down modulation that guides the information processing of lower-level mechanisms (say, in posterior brain structures) and through such top-down modulation, bias the response-selection in a process similar to that of the Cooperative Decision Thesis.

[91] Originally, the algorithms aim to capture the underlying information processing subserving the goal-directed and habitual control of instrumental behavior respectively (Balleine & Dickinson, 1998). Briefly, instrumental behaviors under goal-directed control are flexible in the sense that they are sensitive to the changes in environment and the

implementation in human cognitive systems (Nathanie D. Daw & Dayan, 2014; Dayan & Niv, 2008; Lee, Seo, & Jung, 2012). In the following, I will first review the support MBMF lends to the Hierarchical Structure Thesis. Then, I will discuss briefly how this line of literature also supports both the Embodied Agent Thesis and Cooperative Decision Thesis.

4.1. Hierarchical Structure

MBMF posits hierarchical structure in the information processes that subserve learning and control of instrumental behavior: the higher-level mechanisms implement model-based algorithms, while the lower-level processes implement model-free ones. Moreover, empirical research into the two processes' interaction points to a complex relationship between model-based and model-free processes. It has been suggested that in certain situations, model-based processes can help train the model-free processes by supplementing feedback signals (reward prediction errors). In addition, model-based processes may also incorporate model-free processes as part of their information processing (Nathanie D. Daw & Dayan, 2014, pp. 7–8). However, at the current stage, it is unclear whether the hierarchical relation between the two types of processes goes beyond the weak sense of hierarchy introduced in the last chapter.[92]

The DVs for response options produced by these two types of mechanisms play a central role in this framework. They are referred to as "value representations." An action's value representations are the predicted estimates of the long-run, expected utility the agent will collect following the performance of this particular action in a particular state. For example, consider the problem of deciding which routes to take to go home from the office on a Friday afternoon (illustrated in Figure 5.7). We can think of this problem more abstractly as having various states (i.e., locations on various routes), actions (i.e., going straight, or making left or right turns), transitions (i.e., moving from one state to the other when a particular action is performed), and outcomes (i.e., positive or negative subjective utilities collected at each transition, such as fuel consumption, time cost, scenery, etc.).

In Figure 5.7, the value representation associated with turning left (a particular action) at the intersection (a particular state) is the prediction of the long-run expected utility the agent will collect through a sequence of actions following turning left at this intersection.[93] Because the DVs for response options are value representations, response-selection based on DVs is response-selection based on the highest subjective utility among all options.

---

agent's motivational states. On the other hand, instrumental behaviors under habitual control are inflexible because they are immune to such changes. For example, a rat, which has been trained to press the lever for food pellets, is under goal-directed control if it stops pressing the lever when it detects changes in environmental contingencies (e.g., pressing the lever no longer leads to delivery of food pellets) or when its motivational states change (e.g., it is already fed to satiation). Alternatively, the rat's behavior is under habitual control if it continues to press the lever regardless of these changes.

[92] Neuroscience research has elucidated the neural circuitry involved in model-based and model-free processes. Model-free processes (at least ones involving appetitive reinforcement learning) are supported by dopaminergic neurons in the midbrain that drive plasticity at the target neurons in the striatum; model-based processes implicates the ventral prefrontal areas, the dorsomedial striatum, and possibly the basal lateral nucleus of the amygdala (Nathanie D. Daw & Dayan, 2014). As we can see, it is not obvious that the two types of processes involve hierarchical mental mechanisms arranged in the order of their forward and backward neuronal connections.

[93] That is, it is not just the immediate utility collected, such as the immediate fuel cost, but one that involves getting onto the freeway, making a detour, decreased travel time, ease of driving, etc.

In fact, the MBMF aims to simultaneously solve two problems: learning and decision-making. The agent starts out not knowing the structure of the problems (i.e., the states, actions, transition, and outcomes) and related value representations. The reinforcement learning algorithms then specify how value representations can be acquired through an agent's interactions with the environment. Once value representations are acquired, they can then be used to guide the agent's choice in order to maximize reward over the long run.

Although the original aim of this research framework was to account for overt behaviors (motor responses), it has been extended to account for the control of internal actions, including cognitive responses—for example, the selection of internal representation for maintenance in working memory, or selection of higher-level goals (Dayan, 2012). In other words, both internal and external actions can be driven by either higher-level model-based or lower-level model-free processes.

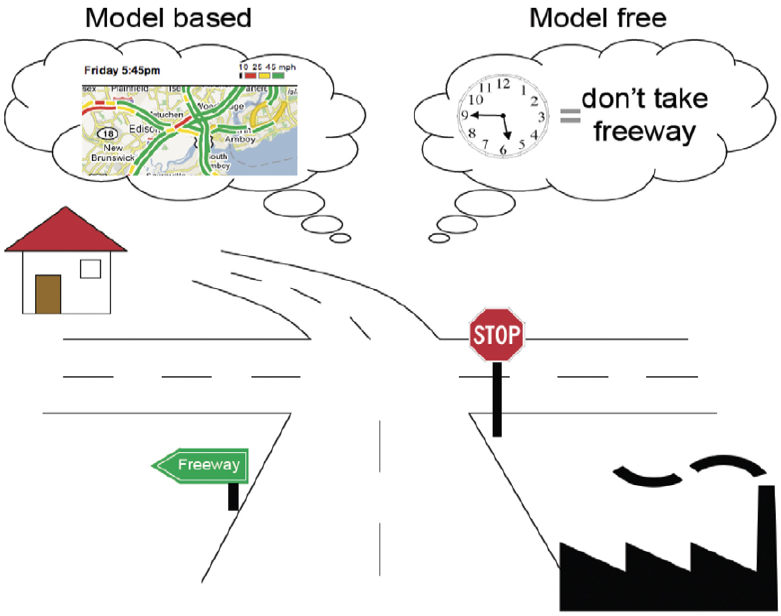

Figure 5.7 Two ways of deciding which routes to take in order to go home from the office on a Friday afternoon. Excerpted from (Dayan & Niv, 2008).

Taken this way, the MBMF is compatible with and can complement the hierarchical models of perception and action-control discussed earlier (as well as the predictive mind framework to be discussed in the following section). This is because the distinction made by the two research frameworks are orthogonal to each other so that, at least conceptually, a particular higher-level or lower-level information process in the hierarchical models of perception and action-control may be a model-based or model-free process, and vice versa. Indeed, computational neuroscientists have started looking at the intersection of the two research frameworks, i.e., the hierarchical model-based and model-free reinforcement learning and control (Botvinick & Weinstein, 2014).

Similarly, the MBMF also supports the systematic difference of higher-level and lower-level processes in terms of flexibility, speed, and capacity. In what follows, I will review the empirical support for these differences.

Flexibility

The MBMF claims that model-based action-evaluating processes are more flexible and capable of producing more accurate DVs in a wider range of contexts. On the other hand, model-free action-

evaluating processes are less flexible and can only produce accurate DVs in those contexts where agents have extensive experiences with (Nathanie D. Daw & Dayan, 2014; Dayan & Niv, 2008; Lee et al., 2012).

What lies at the core of the model-based and model-free processes' difference in flexibility are different ways of updating value representations. Model-based processes produce the value representations through simulation with internal models; hence, the name "model-based" processes. The internal models possess two parts: one is the "transition model" which represents the causal structures of the environment relevant to a given problem (the states, actions, and transitions); the other is the "reward model," which represents the immediate subjective utility (a function of the subject's goal and motivation at the time) to be collected following each individual state or action. The value representations associated with different response options are then calculated through a form of mental simulation. First, a decision tree—with the current state as the root and future states/actions as leaves—is built with the transition model. Following this, a search through the decision tree from the current state to the terminal state is run to accumulate utilities along the way with the help of the reward model.

The model-based processes are flexible because the internal models involved in the processing represent the causal and reward structures explicitly, and they can be updated with information about modifications of the causal structure (e.g., due to environmental changes) or modification of the reward structure (e.g., due to changes in the agent's goals and motivation). For example, the model-based process, illustrated in the "thought bubble" on the left in Figure 5.7, calculates the long-run expected utility of turning left onto the freeway vs. going straight to avoid it respectively, by searching a mental map learned from past experiences. The mental map can be modified with unexpected traffic conditions heard on the radio (road construction, a traffic jam on the freeway, etc.) or a change in goals (going to the gym or home). As a result, the value representation calculated can reflect these changes immediately, and provide the updated and accurate estimates associated with different response options in the current context.

On the other hand, model-free processes do not compute value representation through representing and simulating with internal models—hence the "model-free" algorithm. Instead, the model-free processes merely retrieve the stored value representations of the relevant actions when a decision in a particular state is required. The stored value representations come to track the actual, long-run utility because they are updated with the reward prediction error signals generated during direct interaction with the environment. Reward prediction error signals are the signed difference between the actual long-run utilities collected by the agent following a chosen action, and the current prediction (value representation) associated with the action. Specifically, the value representation of the action chosen will be updated according to the reward prediction error so that it will be brought closer (by some fraction of the prediction error) to the actual reward experienced following the chosen action.[94] In effect, this allows the agent to maintain a value representation based on the *average* long-run utilities obtained following the particular action in a particular state.[95]

---

[94] Also, the value representations of actions not chosen may be updated by so called "fictive reward prediction errors" (Kishida et al., 2015).

[95] The processes involved in the model-free control are in fact more complex than I present here. For example, model-free reinforcement learning faces a critically important issue: the temporal credit assignment problem. That is, how to update a value representation of an action chosen at an earlier point in time with the rewards obtained much later (Richard S. Sutton & Barto, 1998). Some procedure must be done to bridge this temporal differences. Several

Model-free processes are inflexible because the value representations acquired through them are not likely to be accurate across a wide range of contexts. Rather, they are only reliable in contexts in which they have extensive experience. When changes happen in the environment, or to the agent's goals and motivational states, the actual long-run utilities collected following an action will be altered. However, because the value representation can only be slowly updated in model-free processes through reward prediction errors acquired in the actual interaction with the environment, information about these environmental and motivational changes have no direct means to alter the stored value representations. As a result, these value representations cannot immediately reflect these unexpected changes.

For example, as illustrated in the "thought bubble" on the left in Figure 5.7, the value representation of response option at this intersection has been shaped by past experience of driving home from work at 5:45pm, and the experience may have taught the driver that the best way to go home is to avoid the freeway. That is, the predicted expected utility of continuing straight at this intersection is higher than that of turning left onto the freeway. However, the value representations cannot be altered by model-free processes without further learning through direction interaction with the environment. As a result, they cannot be updated immediately, say, upon hearing from radio about an unexpected traffic jam straight ahead, nor can it be updated with the changes of the agent's goal of going to the gym instead of going home.

In short, the MBMF also supports the feature of more flexible higher-level information processes and less flexible lower-level information processes.

Speed

Similarly, in the MBMF framework, the lower-level model-free processes are considered to be faster than the higher-level model-based ones. Because model-free processes generate the control signals—i.e., the value representations—simply by retrieving them from memory, it takes considerably less time compared to the model-based processes. This is because model-based processes generate the value representations through a computationally intensive search through a potentially large space of future states, transitions, and outcomes (Dolan & Dayan, 2013, p. 320).

Capacity

Finally, the model-based and model-free processes also differ in their processing capacity: the model-based processes are capacity-limited compared to the model-free processes. This should not be surprising, because the model-based processes demand enormous computational resources when they search through an exponential combination of future states and outcomes. "Consequently, a model-based agent is confronted with overwhelming computational constraints that in psychological terms reflect the known capacity limitations within attention and working memory" (Dolan & Dayan, 2013, p. 316). In fact, there is empirical research showing that people, when burdened with cognitively demanding tasks that drain their working memory capacity, come to rely more on model-free processes for behavioral control (Otto, Gershman, Markman, & Daw, 2013).

In conclusion: the MBMF also supports the Hierarchical Structure Thesis, which postulates higher-level and lower-level autonomous information processes that systematically differ in their flexibility, speed, and capacity.

---

algorithms have been developed to ameliorate this problem, such as the temporal difference reinforcement learning. I will not go into the details of this point, as it does not influence the main point I need to make in this chapter.

4.2. Embodied Agent

The above characterization of the model-based process, insofar as it involves separate transitional and reward models and calculates value representations as the expected utility of an action, seems like a very "classical" approach to cognition. As a result, it may seem that the MBMF does not fully support the Embodied Agent Thesis. However, the Embodied Agent Thesis is in fact supported by the more biologically-realistic models of the MBMF.

One important qualification to make concerning the seemingly very "classical" features of the model-based algorithm is that the prototypical model-based algorithms discussed above are computationally intractable for many of the real-life problems, because the decision trees grow exponentially over a few steps. Various different methods of approximating the prototypical model-based algorithm have to be implemented in the cognitive system instead. Moreover, these different approximations have "underappreciated consequences for the interdependency of perception and action" (Gershman & Daw, 2012, p. 293). This is because they suggest an embodied vision for the model-based mechanism. Instead of the classical architecture, "organizations for the underlying neural systems [for model-based reinforcement learning and control] involve a richer ensemble of dynamical interactions between perceptual and motivational systems than that which is anticipated by statistical decision theory" (Gershman & Daw, 2012, p. 308).

Additionally, there are also important differences in various versions of model-free algorithms. In fact, the model-based and model-free controls are best seen as forming a spectrum with several intermediate types of control processes, whose characteristics lie between the prototypical model-based and model-free controls (Dolan & Dayan, 2013, p. 320). Moreover, these intermediate, and more realistic, control processes are not likely to operate on detailed, action-neutral representations of the casual and reward structures of the world with algorithms that mirror the sequential stages of perception and action in the classical view. Instead, they rely on partial and action-oriented representations with approximate algorithms that collapse the two stages of state inference and utility calculation in several ways (Gershman & Daw, 2012).

As a result, the MBMF, at this current stage, is at least compatible with the Embodied Agent Thesis, and future research in this area may come to provide strong support for the thesis. Specifically, future research may lend support to both the "universal constitutive dependence on offline embodiment" claim (i.e., that all information processes depend constitutively on sensory and motor mechanisms) and the "representational embodiment" claim (i.e., that some important information processes depend constitutively on modality-specific and action-centric internal representations).

4.3. Cooperative Decision

The traditional way of understanding the MBMF literature does not seem to support the Cooperative Decision Thesis. This is because the traditional formulation often focuses on how the (exactly) two autonomous processes compete with each other for the complete control of behavior. That is, the traditional understanding of MBMF does not consider the way in which numerous heterogeneous information processes (each of which may be placed somewhere on the spectrum of model-based and model-free processes) cooperate to select the best option.

For instance, in the earlier study of control of instrumental behavior, the emphasis is placed on how the goal-directed control supported by model-based processes compete with the habitual control supported by model-free processes working in parallel (Nathaniel D. Daw, Niv, & Dayan, 2005; Dayan & Niv, 2008). An example of this is when our minds wander off while driving on a familiar route; habitual control may sometimes out-compete the goal-directed control, leading us to make a

wrong turn at the intersection that takes us to a location we do not intend to go at that moment. Although most existing literature remains focused on the selection "between" model-based and model-free control, over the last few years research has begun to address the issue of how a "mixed instrumental controller" can flexibly "combine" the outputs (DVs) of both types of control in order to drive instrumental behavior, based on a set of considerations that include accuracy and the cost of information processing (Dolan & Dayan, 2013; Pezzulo, Rigoli, & Chersi, 2013).

Moreover, while most existing research concerning the dynamics of control processes implicitly assumes that the interaction is between one single model-based and one single model-free process, this idealization may not hold in reality. It is plausible that the learning and control of responses are driven collectively by several model-based and/or model-free processes in the human cognitive system. In fact, it has been suggested that these multiple model-based controllers are realized in the human brain (Dolan & Dayan, 2013, p. 320; Doya, Samejima, Katagiri, & Kawato, 2002, p. 1366).

In conclusion, even though the dominant trend in the MBMF literature may not support the Cooperative Decision Thesis, recent research by prominent researchers in this field has begun to formulate and test accounts that support this thesis. Similarly, we can expect to see more MBMF literature supporting the Embodied Agent Thesis in the near future. In short, MBMF has provided good support for the Hierarchical Structure Thesis.

5. The Predictive Mind

A very recent advance in cognitive science is the "predictive mind" framework (Clark, 2013; Hohwy, 2013). Despite significant differences between it and the hierarchical models of perception and action-control discussed above, this framework nonetheless supports the Hierarchical Structure Thesis. In addition, because the predictive mind framework advocates an intimate relation between perception and action, it also strongly supports the Embodied Agent Thesis.

5.1. Hierarchical Structure

The predictive mind framework supports the Hierarchical Structure Thesis because of its strong, albeit idiosyncratic, position on the hierarchical structures of neural mechanisms.

There are two major ways in which the hierarchical structure of the predictive mind framework deviates from those of more traditional hierarchical models. First, it suggests that one of the fundamental tasks of the cognitive system is to predict the inputs it will receive based on its hierarchical, probabilistic models of the causal structure of the world, and to revise its models based on errors of prediction (the mismatch between predicted and actual inputs). As Andy Clark puts it,

> In this paradigm, the brain does not build its current model of distal causes (its model of how the world is) simply by accumulating, from the bottom-up, a mass of low-level cues such as edge-maps and so forth. Instead… the brain tries to predict the current suite of cues from its best models of the possible causes. (Clark, 2013, p. 2)

Implemented in a hierarchy of mechanisms, the higher-level mechanisms constantly predict the inputs to their lower-level ones; in addition, the lower-level mechanisms calculate the mismatch between predicted and actual inputs (i.e., the prediction error), which is then used to update the probabilistic models of their higher-level mechanisms. Importantly, the forward flow of prediction error replaces the forward flow of sensory data (Figure 5.8). For example, the forward connections from V1 (primary visual cortex) to V2 (secondary visual cortex) carry the prediction errors about

the anticipated activities in V1, while the backward connections from V2 to V 1 carry V2's predictions of the anticipated activities in V1.

Moreover, the predictive mind framework also supports the more specific claim of the Hierarchical Structure Thesis, that the higher-level information processes are more flexible than the lower-level ones. However, it does not particularly address the speed and capacity differences between information processes.

In the predictive mind framework, higher-level processes are also more flexible for reasons similar to those discussed in hierarchical models of perception and action-control. Despite the idiosyncratic understanding of the functional significance of forward and backward information, one familiar feature is preserved in the "duplex architecture" of the predictive mind, as Clark (2013) calls it. That is, at each level of the hierarchy, there are the familiar "representation units" that encode the causes of sensory inputs (a world model), in addition to the idiosyncratic "error units" dedicated to computing prediction errors (Figure 5.8). As a result,

> The duplex architecture thus achieves a rather delicate balance between the familiar (there is still a cascade of feature-detection, with potential for selective enhancement, and with increasingly complex features represented by neural populations that are more distant from the sensory peripheries) and the novel (the forward flow of sensory information is now entirely replaced by a forward flow of prediction error). (Clark, 2013, p. 8)

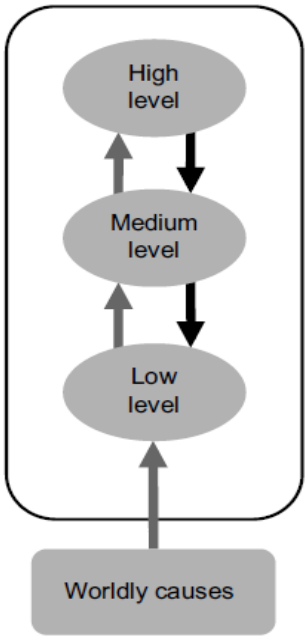

Figure 5.8 A simplified information processing hierarchy in a predictive coding model. Causal regularities of different time scales are processes at different levels (there are more than 3 levels). Information processing at each level influences the others through forward and backward connections. At each level, the gray arrow represents predictive error from the world and from the error units, and the black arrow represents prediction from the representation unit. The aim to minimize prediction errors then drives the perceptual inference at each level simultaneously, forming a structured model of the external environment. Excerpted from (Hohwy, 2013).

In fact, similar to the hierarchical models of perception and action-control, this framework hypothesizes that the lower-level mechanisms process fast and concrete regularities of the world within a "small, detail focused receptive field" (Hohwy, 2013, p. 29), while higher-level ones have wider receptive fields and process more abstract regularities of larger temporal and spatial scales

(Figure 5.8). As a result, higher-level information processes are more flexible because they are modulated by top-down predictions generated by higher-level models, which carry more complex, abstract, and contextual information of the world.[96]

In short, while the predictive mind framework literature presents a very novel picture of how the mind works, it nevertheless supports the general outlook of the Hierarchical Structure Thesis.

5.2. Embodied Agent

The predictive mind framework also supports the Embodied Agent Thesis. This framework suggests an intimate relation between perception and action—they work together to reduce prediction errors. In this framework, perception is a process that attempts to reduce the prediction errors by changing the cognitive system's best hypotheses of the world, and as a consequence, its predictions of incoming sensory inputs. Motor systems suppress the prediction errors, on the other hand, by producing actions that change the world, and as a result, make the incoming sensory inputs conform to the predictions. This intimate relation between perception and action, as well as between neural information process and bodily and environmental feedback, lend support to all three claims of the Embodied Agent Thesis: the "universal constitutive dependence on offline embodiment," "existential constitutive dependence on online embodiment," and "representational embodiment" claims.

Such a tight relation between perception and action can lead to an extreme position, what Clark calls the "desert landscape" version of the theory (Clark, 2013, p. 6). In this version, there is only one unified hierarchical model of the world, rather than separate perceptual and motor models, and prediction errors can act directly as motor commands that drive actions. That is, "the brain has the singular task of predicting sensory input. This means that the generators of motor output simply predict the sensory consequences of anticipated [intended] movements…" (Kiebel, Daunizeau, & Friston, 2008, p. 10). In fact, the desert landscape version of the *predictive mind* will support an extreme version of the Embodied Agent Thesis, according to which information processes depend constitutively, not on perceptual mechanisms and motor mechanisms, but on mechanisms that are perceptual and motor in nature at the same time (i.e., an extreme version of the "universal constitutive dependence on offline embodiment" claim).

In conclusion, the *predictive mind* framework supports the general outlook of the Hierarchical Structure Thesis. It also supports the Embodied Agent Thesis, potentially, in an extremely strong form.[97]

---

[96] We should note that despite the central role of backward prediction in the framework, this top-down information still may or may not be sensitive to the current information processing in a particular lower-level mechanism. As a result, the distinction between information processes that "truly involve" higher-level mechanisms and ones that are "merely modulated" by them still hold in the predictive mind framework (see Section 3.2 in the last chapter). For example, a typical motor response, or an active inference in predictive framework's terminology, is always driven by prediction signals from higher-level mechanisms. However, these prediction signals are not always sensitive to the current information processing at the relevant lower-level mechanisms. That is, the generative model of the world represented in the higher-level mechanism may generate prediction signals that drive a particular active inference before the higher-level mechanism is forced to revise its world model by the prediction error signals it receives from the relevant lower-level mechanisms. In such case, this active inference is a product of a less flexible, lower-level information processes because it is "merely modulated" by higher-level mechanisms.

[97] I also want to note that under certain assumptions predictive mind models and the sequential sampling models of decision-making (to be discussed in Section 7 as support for the Cooperative Decision Thesis) are formally equivalent

6. Dual-Process Theories

The fourth line of evidence that supports the Hierarchical Structure Thesis comes from dual-process theories. Dual-process theories have enjoyed prominence in psychological research recently. They have been proposed to account for very different domains of information processing to those addressed by the literature we have reviewed thus far. Specifically, their domains of information processing include the central cognitions, including but not limited to social cognition, moral judgment and deductive reasoning (J. S. B. Evans & Frankish, 2009; Kahneman, 2011). I am skeptical of dual-process theories' capacity to offer a mechanistic account, as opposed to a phenomenological description, of the human mind. Nevertheless, I include it as an empirical literature that supports HECA in its limiting case.

6.1. Hierarchical Structure

According to one substantial interpretation of *dual-process theories* (Samuels, 2009), two types of systems, type 1 and type 2, are postulated to subserve two corresponding types of processes. Additionally, there is a hierarchical relation, in the weak sense at least, between type-1 and type-2 systems.[98] According to this theory (Figure 4.3 in p.70), one kind of the lower-level type-1 (pre-attentive) process supplies content and introduce knowledge to the higher-level type-2 (analytic) processes. Type-2 processes, by inhibiting, endorsing, or replacing type-1 responses, can intervene on the workings of type-1 processes (pre-attentive processes and autonomous process).

In the following, I will briefly discuss the support dual-process theories lend to the Hierarchical Structure Thesis, as well as the systematic differences between higher-level and lower-level information processes.

Flexibility

Information processing in type-1 systems is often claimed to be reflexive, intuitive, associative, and heuristic, while information processing in type-2 systems is considered reflective, analytic, rule-based, and systematic (J. S. B. T. Evans, 2008). Although the correct functional characterizations and causal mechanisms of both systems remain hotly debated (J. S. B. T. Evans, 2008, p. 261; Frankish, 2010, p. 921), it is clear that most researchers consider information processes involving type-2 systems to be superior and more flexible to those involving only type-1 systems.[99, 100] For instance, it is believed that "the former is often associated with normatively correct responding and

(Summerfield & de Lange, 2014, p. 752). As a result, we may also think that the predictive mind framework is supportive of the Cooperative Decision Thesis.

[98] However, because research into the neural substrates of type 1 and type 2 systems is still in its infancy, it is unclear whether they can be arranged into an information-processing hierarchy in the strong sense, according to their forward and backward neuronal projections (Lieberman, 2007). For now, we can take the hierarchical structure in the strong sense to be one possible empirical implementation of dual-process theories.

[99] It remains unclear what is behind the type-2 processes' flexibility. It may be due to the rule-based information processing and/or the working memory mechanisms involved in type-2 processes.

[100] Note that this claim is compatible with what many researcher now accept: type-2 processes may deliver normatively incorrect response (Frankish, 2010), and that type-1 processes may often generate the right responses, especially in expert decision-making (J. S. B. T. Evans, 2008, p. 267). Note that my main argument will not require the stronger claim that higher-level processes will always produce normatively correct responses, but the weaker claim that higher-level processes will tend to produce normatively correct responses in a wider range of contexts.

the latter with cognitive biases" in many areas of central cognition, including deductive reasoning, probability judgment, and social cognition (J. S. B. T. Evans, 2008, p. 267).[101]

Speed

The feature of speed differences between higher- and lower-level information processes is perhaps made most explicit in the dual-process theories (J. S. B. T. Evans, 2008). In *Thinking, Fast and Slow*, Daniel Kahneman considers the speed of information processing one of the most defining features of the type-1 and type-2 systems: "I describe mental life by the metaphor of two agents, called System 1 and System 2, which respectively produce fast and slow thinking" (Kahneman, 2011, p. 13). For example, when presented with a moral judgment task, experimental subjects usually quickly experience an intuitive "gut feeling" about the situation (generated by a type-1 system) before the slow deliberative reasoning processing (subserved by a type-2 system) kicks in to generate a more thoughtful judgment (Haidt, 2001).

Capacity

Similarly, in dual-process theories, type-2 systems are usually associated with the attribute of low capacity and type-1 systems with the attribute of high capacity. According to one proposal, the difference in capacity between the two systems originates from the fact that type-2, but not type-1, systems processes "require access to a single, capacity-limited central working memory resource" (J. S. B. T. Evans, 2008, p. 270). As a result, there can be many more type-1 processes than type-2 processes operating at the same time.

In short, the characterization of type-1 and type-2 systems in dual-process theories supports the Hierarchical Structure Thesis in its limiting case—that is, when there are only two levels of cognitive mechanisms.

6.2. Embodied Agent and Cooperative Decision

In the following, I will suggest that the dual-process theories can be interpreted to support the Embodied Agent and Cooperative Decision Theses in their limiting case as well. In making this case, my intention is not to use dual-process theories to lend strong support for HECA; rather, I hope to show that the image of mind HECA puts forward is so prevalent in contemporary cognitive science that its faint outlines can be found in dual-process theories.

First, dual-process theories are committed to a weak version of the Embodied Agent Thesis. There is a clear commitment to the autonomy of information processes in dual-process theories, according to which type-1 processes (responsible for "automatic" cognition) can often generate responses without involving type-2 processes. Type-1 reasoning processes can produce judgments directly that may or may not conflict with the judgments produced by type-2 reasoning processes in the parallel-competitive model of dual-process theories (J. S. B. T. Evans, 2009, p. 43). Either one or the other type of processes may take control of the behaviors. For example, people can sometimes act out on an automatic racist judgment, contrary to their deliberative judgment of egalitarianism. In fact, many prominent researchers, taking stock of the massive literature on automaticity, go as far as claiming that the majority of our mental activities are the result of such "automatic" cognition

[101] What counts as a normatively correct response in a particular situation and what sort of normative standard is appropriate in evaluating a particular judgment are issues under debates (Gigerenzer, 2006, p. 119), but the main argument of this chapter does not depend on settling these issues in any particular way.

(Bargh & Chartrand, 1999; Kahneman, 2011).[102] Moreover, the characterization of type-1 responses as "hot" gut feelings can suggest that (at least some) type-1 processes may be sensorimotor and bodily in nature.

Moreover, dual-process theories may also be committed to a weak version of the Cooperative Decision thesis. It is true that the literature usually emphasizes the conflict between a single type-1 and a single type-2 process; the conflict, for example, between the "hot" gut feelings generated by type-1 processes and the "cold" deliberative judgment generated by type-2 processes. However, it is worth noting that phenomenology suggests there can be conflicts not only between different types of processes, but between processes of the same type. For example, an agent can experience two conflicting intuitions concerning appropriate etiquette in a particular social situation or two conflicting deliberative judgments about their moral duty. In some cases, an agent can even experience several conflicting intuitions as well as deliberative judgments all at the same time. These conflicting type-1 and type-2 information processes may cooperate with each other in order to select a response in the way the Cooperative Decision Thesis describes.

In conclusion, the dual-process theory literature, while very different in its domains of application compared to other literature reviewed above, nevertheless supports HECA's core features in their limiting cases.

In the last four sections, I have primarily focused on empirical support for the Hierarchical Structure Thesis. Specifically, this support was derived from hierarchical models of perception and action-control, model-based and model-free reinforcement learning and control, the predictive mind, and dual-process theories. These accounts all reflect a general tendency in contemporary cognitive neuroscience to understand human cognitive systems as consisting of hierarchies of mental mechanisms, which subserve autonomous information processes that specify and evaluate actions with varying flexibility, speed, and capacity. In the next section, I change direction to look at an important development in neuroscience that supports the Cooperative Decision Thesis.

7. Sequential Sampling Models of Decision-Making

I've shown in the previous sections that important features of HECA are supported empirically by various large-scope models of cognition. In this section, I focus on empirical support for the Cooperative Decision Thesis derived from an important class of mathematical and computational models of decision-making: the sequential sampling models (SSMs) (Busemeyer & Johnson, 2004; Ratcliff & McKoon, 2008; Usher & McClelland, 2001).

The SSMs were first developed as mathematical models that account for decision-making in the more "central" cognitive domains. However, they have been extended to response-selection in the sensorimotor domains (Cisek & Pastor-Bernier, 2014, p. 7). Moreover, over the last 20 years, empirical data in the neuroscience of decision-making or neuroeconomics concerning the neural basis of decision-making has corroborated the mathematical models (Gold & Shadlen, 2007) and led to a series of neurocomputational models that attempt to offer biologically-inspired mechanistic account of decision-making (Usher & McClelland, 2001; Wang, 2002). Importantly, these models, although differing in their details, all assume that response-selection is based on the dynamic accumulation of DVs for alternative options until the selection of the option whose DVs exceed the

[102] Automaticity is a complicated concept in psychological science, the definition of which remains controversial and problematic (Moors & De Houwer, 2006). However, we will not be concerned with the nature of automatic and controlled processes *per se* in this chapter. So, I will leave this issue aside for now.

threshold first. Finally, all the literature reviewed in the previous sections are compatible with the sequential sampling models of decision-making, although not all of them commit themselves explicitly to these models.

In the following, I will review the core mechanistic components and their operation in the SSMs, as well as their empirical support. I will then briefly describe how they implement the sequential probability ratio test and support the Cooperative Decision Thesis.

Here, I illustrate the SSMs' basic components and their dynamics with one particular model: the *non-linear leaky competing accumulator* (non-linear LCA) model (Bogacz, Usher, Zhang, & McClelland, 2007; Usher & McClelland, 2001). A non-linear LCA architecture for two competing alternatives is shown in Figure 5.9. It includes two accumulator units (top circles in Figure 5.9) corresponding to neuronal populations that integrate DVs (i.e., accumulate the neuronal activities in time) for two alternatives, respectively. Two units at the bottom correspond to neuronal populations that provide evidence to each alternative.[103] There are mutual inhibitory connections (represented in red) between the two accumulator units, and the level of inhibition depends on the accumulator's current activities (i.e., the stronger an accumulator's current activities is, the stronger can it inhibit the other accumulator). Moreover, the accumulators are modeled as leaky integrators (hence, the water drops on either side of the accumulators) so that their activities decay at a specific rate. It is a non-linear model because the activity at an accumulator is transformed by a non-linear function so that it will remain within a range that is biologically plausible (i.e., it cannot be negative and cannot exceed a certain level, just like a real neuronal population).

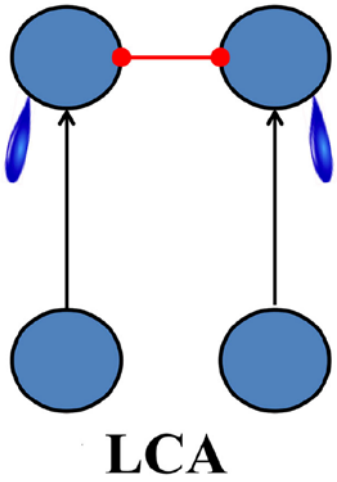

Figure 5.9 The leaky competing accumulator model (Usher & McClelland, 2001). Reprinted from (Tsetsos, Gao, McClelland, & Usher, 2012).

Here is an example of the dynamics of the non-linear LCA model for a simple perceptual decision task. Perceptual decision tasks are simple decisions that often involve identifying and distinguishing certain environmental features through perception, and indicating the presence or absence of the feature with simple motor movement (such as pressing a button). In a paradigmatic random dot motion discrimination task (as discussed in Chapter 4's Section 3.3), the subject (a computational model here) is presented with visual stimuli of a population of dots moving largely randomly, but

[103] The Non-linear LCA model, like other SSMs, is an idealized model of our cognitive systems. For example, in reality, there would likely to be more than one neuronal populations providing evidence for each alternative in the cognitive systems.

with a small trend of moving in one of the two possible (and opposing) directions. The subject needs to identify which one of the two possible directions the dots are moving toward, and indicate the answer. The dynamics of the two accumulators in this model is illustrated in Figure 5.10. When the visual stimuli are presented, the neuronal populations that detect visual movements start producing DVs in the form of neuronal activations. At the same time, the two accumulators also start integrating these incoming activations, which can be represented by the two initial rising trajectories in the Figure 5.10. Because one accumulator (whose activity trajectory is red) receives more evidence than the other, it exerts a stronger inhibition on the other accumulator. This results in the other accumulator's activity decreasing slowly until it hits zero. The stronger accumulator continues to increase its activity level due to more incoming evidence and the diminishing inhibition from the other accumulator. The activation accumulation continues until the activity reaches the threshold for decision at time t, at which time the perceptual decision is made concerning which direction the random dots are moving towards.

Moreover, abundant behavioral and neuroscientific data support the SSMs. First, it has been established that the neural mechanisms of perceptual decision-making exhibit features that are characteristic of the SSMs.[104] For example, we can identify the neural correlates of several important parameters of the models and their dynamical changes,[105] and the models can predict relatively successfully the behavioral data, such as the reaction time and error rates of perceptual decision tasks (Gold & Shadlen, 2007; Shadlen & Kiani, 2013).

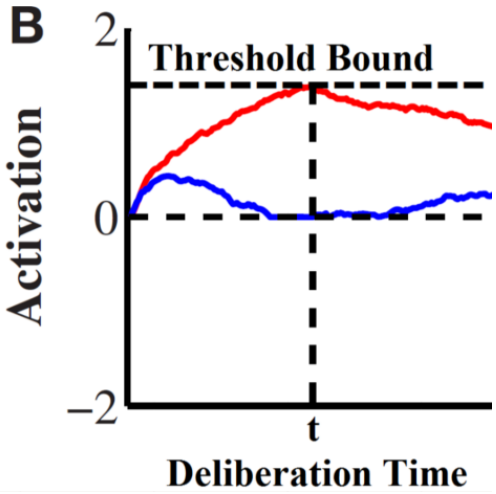

Figure 5.10 Activities of two accumulators in a *non-linear LCA* model. Adapted from (Tsetsos et al., 2012).

Second, there is evidence that value-based decisions, i.e., decisions based primarily on the expected utilities of options, are also subserved by sequential sampling processes. Preliminary evidence from

[104] Most research on perceptual decision-making is done on the vision, but there is increasing research on decision-making related to auditory, tactile, olfactory, and other senses (Shadlen & Kiani, 2013, p. 797).

[105] There is evidence showing that, in the paradigmatic random dot motion discrimination task, the brain area MT, associated with detecting visual movement, is responsible for generating DVs and the brain area LIP, an area midway of the sensorimotor loops between MT and FEF (an area associated with controlling eye movement), is responsible for accumulating the DVs. The dynamics of the activities at the areas of MT and LIP can also be well predicted by relevant sequential sampling models. For more detailed review, see (Gold & Shadlen, 2007, p. 545).

neuroscience exists (Gold & Shadlen, 2007, p. 560; Platt & Glimcher, 1999). However, the stronger evidence comes from purely behavioral studies: researchers have shown that SSMs of value-based decision-making can explain several significant anomalies of decision-making that cannot be accounted for elegantly by the traditional expected utility theory (Busemeyer & Johnson, 2004; Oppenheimer & Kelso, 2015). More importantly, these behavioral studies in the SSMs illustrate to us how interesting higher-level (in the part-whole sense) phenomena of human decision-making emerge out of interactions of lower-level component processes.[106] As a result, they provide explanations of how (more or less) intelligent behaviors and the relevant high-level competences (in the part-whole sense) emerge from collaborative interactions of information processes.

Finally, there is also preliminary evidence that the SSMs describe the neural basis of more central cognitive processes, such as the selection of higher-level goals and abstract rules, routing of information, long-term memory retrieval, and probabilistic reasoning (Lee et al., 2012, p. 294; Shadlen & Kiani, 2013, p. 799).[107, 108]

Now we turn to the relation between SSMs and the sequential probability ratio test. Although different SSMs vary in their details, an important subset of them (including non-linear LCA) can all be parameterized to implement the same optimality model, the sequential probability ratio test (Bogacz, 2007, p. 118).[109] We can get a rough and qualitative idea of how the non-linear LCA

---

[106] For examples, anomalies of decision-making discussed in Section 3.3 in Chapter 4 can be elegantly explained by some SSMs (Busemeyer & Johnson, 2004, 2008).

[107] The selection of higher-level goals involves decision about goals that are not directly linked to specific motor actions, such as choosing to have cake or fruit for dessert. The selection of abstract rule involves choice of what to do when a specified antecedent is satisfied, e.g., when the predators show up, one will fight or flee. Routing of information involves determining whether a piece of information is relevant and allowing it to enter a particular (downstream) information mechanism, e.g., updating of working memory and allocation of attention (Shadlen & Kiani, 2013, p. 800). All of the above are types of cognitive action.

[108] There is some reservation whether the SSMs could adequately address value-based decision-making, and specifically, value-based decision's special feature of "unpredictability" (Gold & Shadlen, 2007, p. 560). Roughly, value-based decision-making seems to involve a process similar to a random number generator. Given a binary decision with one choice better than the other with a subjective probability 0.7, humans (and other animals) seem to decide what to do probabilistically as if flipping a biased coin that will come up with the first choice with probability 0.7 (Shadlen & Kiani, 2013, p. 799). What is at stake here is whether this feature of unpredictability "reflects a decision mechanism that explicitly generate randomness" (Gold & Shadlen, 2007, p. 561) and as a result cannot be accounted for by the deterministic sequential sampling models, or reflects simply a deterministic decision mechanism that is unfortunately faced with unavoidable noisy inputs. However it turns out, this debate shall not impact on Cooperative Decision Thesis's core shape.

[109] All sequential sampling models can be categorized into two types: the diffusion models and the race models. In the race models, evidence supporting each alternative is accumulated independently until one of them reaches a fixed threshold. In the diffusion models, evidence supporting for each alternative is accumulated until the *difference* between that of the winning alternative and that of the losing alternative exceeds the threshold (Bogacz, 2007, p. 118). Diffusion models have been shown to be more optimal than the race models—diffusion models take shorter time to make a decision with the same reliability, because their decisions are sensitive to additional pieces of information, that is, the *differences* between evidence for alternative options (Bogacz, 2007, p. 118). In addition, it has been shown that all of the major diffusion models for two alternative choices (Shadlen & Newsome, 2001; Usher & McClelland, 2001; Wang, 2002) are computationally equivalent to each other under parameter values that optimize their performance; they also all implement the *sequential probability ratio test*, which will be discussed briefly in this section (Bogacz, 2007, p. 119).

models introduced above implement the sequential probability ratio test by comparing Figure 5.10 and Figure 4.2 (p.67). Roughly, non-linear LCA models use two competing non-linear accumulators with mutual inhibition to implement the accumulation of linear DVs in the sequential probability ratio test. The positive and negative criteria in SPRT are implemented by the threshold of each accumulator respectively. In other words, we can see these SSMs as mechanistic accounts at Marr's algorithmic or implementational levels of analysis that correspond to the dynamical vision of optimal decision-making at the computational level of analysis, the sequential probability ratio test.[110]

In short, the SSMs are both biologically plausible and empirically well-supported. They also have the potential to account for all types of decision-making, be it perceptual or value-based. Most Importantly, the SSMs, by implementing the sequential probability ratio test, lend strong support to the Cooperative Decision Thesis.

8. Conclusion: Progress and Remaining Problems

> In our brains there is a cobbled-together collection of specialist brain circuits, which, thanks to a family of habits inculcated partly by culture and partly by individual self-exploration, conspire together to produce a more or less orderly, more or less effective, more or less well-designed virtual machine, the *Joycean machine*. By yoking these independently evolved specialist organs together in common cause, and thereby giving their union vastly enhanced powers, this virtual machine, this software of the brain, performs a sort of internal political miracle: It creates a *virtual* captain of the crew, without elevating any one of them to long-term dictatorial power. Who's in charge? First one coalition and then another, shifting in ways that are not chaotic thanks to good meta-habits that tend to entrain coherent, purposeful sequences rather than an interminable helter-skelter power grab. (Dennett, 1991, p. 228)

By updating Dennett's Pandemonium architecture with recent advancements in cognitive neuroscience, I hope that I have made anti-Cartesian psychology less miraculous and more mechanistically grounded. HECA incorporates the developments of autonomous embodied information processes, hierarchically-structured neural mechanisms, and dynamical cooperative response-selection that are supported by prominent large-scope empirical theories of cognition. The six lines of computational and neuroscientific models reviewed above represent a wide spectrum of theories and phenomena. They may not support the specific details of HECA to the same degree, and some of them may even contradict each other.[111] However, looking from a higher

---

Additionally, although all major diffusion models can be parameterized to perform the same computation, each model still differs from each other when the parameters are set to its own typical values. *Non-linear LCA* stands out here for its neurobiological realism as its parameters are constrained by biologically realistic assumptions (e.g., the non-linear feature of the integrators of evidence mimics the features of a real neuronal population) (Bogacz, Usher, Zhang, & McClelland, 2007). There is also some evidence suggesting that the response competition it implements (i.e., mutual inhibitory connections between integrators) is more supported empirically than the other models. For more discussion, see (Teodorescu & Usher, 2013). For these reasons, I base my discussion of the *DDM* on the *Non-linear LCA*.

[110] Different SSMs describe decision processes at different level of analysis. Some are psychological models of behaviors at the more algorithmic level, while other models at the more implementational level aim to capture details of neural circuits of decision-making (Bogacz, 2007, p. 118).

[111] For example, the affordance competition model, the hierarchical action-perception models, the model-based/free learning and control theories, and the sequential sampling models of decision-making can be formulated to be

vantage point, and abstracting away from the particularity of individual theories, these major lines of literature jointly show us a robust trend in cognitive neuroscience that is captured by HECA.

According to HECA, human cognitive systems are composed of embodied autonomous information processes. They are subserved by hierarchical sensorimotor mechanisms, a multiplicity of action-specifying and action-evaluating mechanisms of different types, situated at various levels of hierarchies. The higher-level information processes, enabled by higher-level mechanisms, tend to be more flexible, but slower in speed and limited in capacity, and vice versa for the lower-level information processes. Moreover, these autonomous information processes do not take orders from executive systems, such as Fodorian central systems. Instead, they cooperate with each other to select the optimal responses across the cognitive system: action-specifying information processes specify different response options, and action-evaluating information processes evaluative these options for their optimality for the subject in the current context, based on different standards and information. The dynamical cooperative selection process is ecologically rational, because it is based on the accumulation of evaluations or evidence toward a decision threshold that is sensitive to the demands of time, resources, and accuracy in particular contexts. Finally, responses selected at one mechanism (through subserving other action-specifying and action-evaluating information processes) will influence other response-selections at the other mechanisms, and do so reciprocally as well. Ultimately, intelligent thoughts and behaviors emerge from this distributed consensus-building processes.

Although HECA remains a mechanistic sketch (Craver, 2007) with many relevant details to be filled in and challenges to be overcome, it contains important computational and mechanistic details and allows us to move forward theoretically on the control problem. Again, control problem is the problem of how the diverse neural mechanisms coordinate with each other in the production of intelligent behaviors. I've touched on the problem of architecture briefly in Chapter 1. I will now focus on the problems of coherence and intelligence. In the rest of this chapter, I will evaluate the progress HECA makes in addressing the control problem as an embodied cognitive science approach. As we will see, HECA also raises further challenges for control that an embodied cognitive architecture needs to address. I will briefly motivate and elaborate them, and suggest specific questions that can be addressed in order to overcome these challenges.

8.1. Problem of Coherence

The problem of coherence concerns the challenge of making coherent control decisions in order to generate large-scale, goal-directed behaviors in a highly fragmented cognitive architecture, such as HECA. As I have discussed in Chapters 1 and 2, classical cognitive science's vision of central executive control systems—which rely on rich message-passing and detailed world models—is neither biologically plausible nor empirically supported. However, embodied cognitive science's vision of distributed neural control structures with simple message-passing and proprietary models, in turn, seems insufficient to produce complex and large-scale coherent behavior.

As I discussed in Chapter 4's Section 3.2, neural control structures are simple message-passing neural mechanisms which function to modulate other neural mechanisms. They integrate relevant information in order to coordinate the "bag of tricks" into exhibiting coherent and goal-directed behaviors. The distributed neural control structures seem insufficient for solving the problem of

---

compatible with each other, as in the new framework of hierarchical model-based/free learning and control theories (Balleine, Dezfouli, Ito, & Doya, 2015; Botvinick & Weinstein, 2014). Together they form a strong support for HECA. On the other hand, it is not obvious that the dual-process theories and the predictive coding framework can be made compatible with the other four theories and with each other without some serious work.

coherence: it is unclear that the set of distributed neural control structures, each of which has only partial information about the world, can coordinate with each other and with other neural mechanisms to make a coherent set of control decisions. To put it another way: if neural mechanisms need to be coordinated by a neural control structure in order to behave coherently, doesn't this distributed set of neural control structures, each of which may promote different (and potentially conflicting) ways of achieving overall coherence, need to be coordinated in order to perform coherently as well? Consequently, questions remain concerning what would, in turn, coordinate these now large and fragmented sets of control structures, and what would resolve their conflicts for the better rather than the worse. It appears that addressing the problem of coherence by positing neural control structures creates a meta-problem.

HECA makes some progress on the problem of coherence. Specifically, it makes progress by incorporating a "hierarchy" of neural control structures. Instead of imposing one level of first-order control structures, HECA posits multiple levels, each of which controls the control structures at its lower level. Thus, when it comes to answering the question of what is responsible for coordinating the large and fragmented set of first-order control structures, the answer HECA provides is *second-order control structures.* These slightly smaller number of second-order control structures are, in turn, coordinated by third-order control structures, and so on. As we move up in the hierarchy, the numbers of control structures reduce and, potentially, the problem of coherence becomes less serious. This is because it is easier for 10 than it is for, say, 100 control structures to coordinate with each other without a higher-order control structure modulating them. In other words, the problem of coherence is mitigated because no neural control structures are necessary to coordinate the neural control structures at the highest level.

I believe that HECA's solution is only partially successful because I don't think a large and highly complex cognitive system can produce complex and coherent behaviors without some kind of central control structure that can integrate a large scope of information and issue coherent control commands. In the next chapter, I will review empirical literature that suggests that the basal ganglia play an important role in addressing the problem of coherence. The BG seem to be evolution's solution to the dilemma of classical cognitive science solution of central systems vs. embodied cognitive science solution of distributed control structures. The groups of subcortical structures can play the role of central selection mechanisms (Redgrave, Prescott, & Gurney, 1999): the basal ganglia's strong, reciprocal connection to most areas in our central nervous system puts it in the right place for the centralized control function. However, its design respects the vision of simple message-passing and highly proprietary models; it relies on an extremely simple model for its functioning and communicates with other neural structures with simple signals of inhibition and activation. In the next chapter, I will discuss in detail these features of the basal ganglia and how they help create coherent behaviors.

8.2. Problem of Intelligence

The problem of intelligence concerns how intelligent control decisions and their relevant competencies emerge from the interactions of less intelligent neural components. The sequential probability ratio test and the sequential sampling models help us understand how intelligence can emerge from an epistemic cooperative process of accumulating DVs from heterogenous information processes until a decision threshold.

Despite this progress, we still fall far short of a complete understanding of how human intelligence emerges. One reason for this is that the sequential probability ratio test and sequential sampling models remain too idealized and do not take into account the fact that different information processes in HECA have different flexibility, speed, and capacity. To illustrate this, I will introduce

three related challenges for managing dynamics, capacity, and accuracy that can prevent the more optimal response options from being selected in the decision process.

First, the challenge of managing dynamics is a challenge for the decision-making process to overcome its tendency to select the speedy but less optimal response option so that it can reliably select the optimal one. In HECA, higher-level information processes, due to their flexibility, are more likely to specify more optimal options or produce more accurate DVs in a particular context. In contrast, lower-level information processes are more likely to specify less optimal options or produce less accurate DVs. This is true especially when the context is novel or complex. This is because (1) lower-level processes, due to their inflexibility, need to be trained to produce more optimal or accurate outputs in a novel context, and (2) the complex contexts may require higher-level mechanisms' involvement (e.g., to provide richer contextual information and complicated information processing) for an information process to produce optimal or accurate outputs. However, the trade-off of flexibility is speed. Because lower-level processes are fast, when confronted with situations, they may specify less optimal options and contribute inaccurate supporting DVs early enough to lead to the selection of the less optimal options.

In short, given the Cooperative Decision Thesis, it is unclear how an optimal response, when it requires higher-level information processes to specify or provide DVs in support of its selection, can be selected reliably. For example, Figure 5.10 can represent a selection between a less optimal response (represented in red) and a more optimal higher-level one (represented in blue). The more optimal response fails to be selected because the less optimal one receives more (positive) DVs from faster lower-level action-evaluating processes, and it reaches the threshold first before the more optimal response has the opportunity to accumulate enough DVs.

As a result, to make response-selection more reliable, the challenge of managing dynamics needs to be dealt with. For example, a dynamics-managing control process may prevent the early and unreliable selection of less optimal responses by setting a higher decision threshold. This will require an option to have more accumulated DVs, which will take longer to accumulate, to be selected. As a result, the higher-level action-specifying and action-evaluating information processes will have time to catch up, and specify or provide DVs for the more optimal response option to be selected.

However, setting a high threshold uniformly will slow down the response-selection uniformly as well, which may create a very optimal cognitive system that does not do much at any given time (i.e., a perfectionist cognitive system). This kind of system is generally a bad idea for organisms living in the real world. A more adaptive solution, then, is to adjust the threshold based on an assessment or prediction of the accuracy of the DVs of a specific information process, or that of the accumulated DVs as a whole. For example, a dynamics-managing control process may increase the threshold to require more DVs to select a response option when it predicts that the current accumulated DVs are of low accuracy and may lead to unreliable decisions. In short, one way to adaptively deal with the challenge of managing dynamics is to adjust the decision threshold based on the assessment or prediction of the accuracy of DVs.

Second, the challenge of managing capacity is a challenge to selectively involve information processes (especially, the more capacity-limited higher-level ones) in a response-selection when doing so would likely be worthwhile. Information processes are limited resources for any realistic cognitive system. In addition, there is a more limited capacity for higher-level information processes than for the lower-level ones in HECA. As a result, higher-level information processes cannot participate in every response-selection. At the same time, involving more action-evaluating information processes—in particular, the higher-level ones—will generally increase the reliability

of response-selection. As a result, information processes in general, and higher-level information processes in particular, should participate selectively in response-selection processes that would benefit greatly from their involvement—for example, decisions that are of high-stake (i.e., where the benefit or cost of making the wrong decision is large) and decisions that are difficult to make (i.e., those in novel or complex contexts). Only this way can the overall payoff of interacting with the world be maximized.

The Cooperative Decision Thesis does not specify how information processes are selectively recruited in order to deal with the challenge of managing capacity. Some capacity-managing control processes need to identify response-selections that can benefit more from the involvement of more information processes (higher-level ones in particular) and recruit them selectively.

Again, one way for the capacity-managing control processes to identify selection processes that can benefit greatly from the involvement of more information processes is to detect or predict the accuracy of DVs. When the accuracy of DVs is low, the response-selection is less reliable. The recruitment of higher-level information processes, which can contribute more accurate DVs, can lead to an increase in the reliability of response-selection and the selection of more optimal responses. On the other hand, if the accuracy of DVs is already high, the recruitment of more information processes may not increase the reliability of response-selection much and is less likely to make a difference to the response selected.

So far, I have discussed the challenges of managing dynamics and capacity. As we've seen, detecting and predicting the accuracy of DVs are useful strategies (among others) employed by dynamics-managing and capacity-managing control processes.

Finally, the challenge of managing accuracy is the challenge of selecting the more optimal response based on the accumulation of DVs generated by information processes of variable reliability. Again, higher-level information processes are more flexible and likely to contribute more accurate DVs in a particular context, while lower-level information processes are less flexible and likely to contribute less accurate DVs. This creates a problem for response-selection because adding up DVs to a decision threshold will not always result in the optimal decision, when (1) the DVs contributed by information process are not always accurate, and (2) there are more DVs contributed by large-capacity, lower-level information processes that are likely to be unreliable. For example, if we want to know the answer to a theoretical physics problem, it will not be helpful to poll a group of 90 high school students and 10 theoretical physicists. As a result, given the Cooperative Decision Thesis, it is again unclear how optimal response can be selected reliably.

One way to overcome this challenge is for some accuracy-managing control processes to assess the accuracy of the DVs of individual processes and weight the DVs of each individual process according to how likely they will be accurate before adding them up. Alternatively, another way to overcome this challenge is for the control processes to assess the overall accuracy of the accumulated DVs, and increase the threshold if the overall accuracy is likely too low to select the optimal option reliably.

In conclusion, HECA has made some progress on the problem of intelligence. However, it has not overcome the challenges of managing dynamics, capacity, and accuracy, and these threaten to prevent the reliable selection of optimal response options. In the cognitive system, different control processes for managing dynamics, capacity, and accuracy may utilize different strategies to overcome these challenges. Moreover, I have shown that at least one of the common solutions to these challenges is to try to identify the accuracy of individual or accumulated DVs, then use this information to increase the decision threshold, recruit more (accurate) DVs, or weight the DVs accordingly.

We therefore need a richer account of how various control processes facilitate this cooperative decision-making in order to manage the aggregation of DVs and increase the reliability of selecting the optimal response. In Chapter 6 and Chapter 7, I will draw on computational and neuroscientific models of the basal ganglia and the literature of social choice theory. I will show how the principle of "neurodemocracy" implemented in the basal ganglia may help tackle these challenges and explain the emergence of intelligence from less intelligent component processes.

To conclude this chapter, HECA can be seen as an empirically-updated development of the society of mind model. HECA's three developments, the Embodied Agent, Hierarchical Structure, and Cooperative Decision Theses, address some unresolved problems as to how intelligent and coherent behaviors can emerge out of the decentralized coordination of a diverse network of neural mechanisms. However, the problems of intelligence and coherence remain challenging. Our brains utilize many other "tricks" we may not presently possess knowledge of in order to overcome the challenges related to these problems. I have raised some of these challenges in this chapter. In the following two chapters, I will attempt to make some progress with the help of cognitive neuroscience and formal epistemology.

Part III

Neurodemocracy

6

Neurodemocracy: Basal Ganglia as the Central Controller for the Embodied Mind

1. Introduction

In the last chapter, I discussed HECA, an updated society of mind account, as the embodied cognitive science solution to the control problem. As we've seen, although HECA makes good progress on the problems of intelligence and coherence, there are challenges it nevertheless fails to address. In this chapter, I will introduce the account of neurodemocracy and argue that it is a conceptually promising and empirically supported account, which will also make some headway on the control problem.

As discussed in Chapter 1, traditional solutions to the control problem usually follow one of two approaches. First, the classical cognitive science approach assumes that central control mechanisms, such as the Fodorian central systems, exert executive control over other neural mechanisms; they do so with the help of detailed internal models of the world, and communicate with other mechanisms via rich messages with detailed representational content (Fodor, 1983). Second, the embodied cognitive science approach usually assumes that the control problem is solved through the self-organization of distributed controllers without centralized control; in addition, the control commands consist in simple messages whose content rarely go beyond activation and inhibition (Clark, 1998). So far, neither of them have provided satisfactory answers to the control problem. However, it is not difficult to see the other conceptually possible approaches available. For example, there is an alternative approach to central control mechanisms that coordinate with simple messages.

In this chapter, I will articulate and defend neurodemocracy, a hybrid account that combines this alternative approach with the embodied cognitive science approach. That is, it is an architecture that utilizes both centralized and distributed controllers with a simple message-passing strategy to solve the control problem. I will take HECA and its account of distributed controllers (i.e., the hierarchical neural control structures) for granted, because I have already defended it in Chapter 4 and Chapter 5. Instead, I will focus on articulating the function of the central controllers and their interactions with the distributed controllers. Based on empirical and computational research in the basal ganglia (BG), I will show that the BG, a group of subcortical structures that play an important role in decision-making and learning, function as central controllers, coordinating and working with distributed controllers in the other brain areas.

First, however, I wish to use the metaphor of a democratic society to initially motivate and outline my neurodemocracy account, before articulating it in more detail. In this society of mind, routine control decisions are delegated to "local governments" (the distributed controllers) to settle locally. Yet, when novel problems arise that cannot be settled this way, the "central election commission" (the BG) steps in, and calls for national elections to elicit inputs from the wider population (action-evaluating information processes distributed across the brain). Then, the BG determine the result on the basis of the votes (DVs) cast by action-evaluating processes. It is important to note that many elections can be held by the BG at the same time for different decisions; as a result, these elections can inform each other and lead to more coherence in their results. Moreover, through these democratic procedures, the BG also enhance the intelligence of the mind's final "collective" decisions.

However, instead of implementing the winning actions, the BG simply delegates implementation to distributed controllers. Nonetheless, the BG do enforce accountability: an information process whose DVs lead to bad decisions will be given less weight at the next election (or more weight if they are conducive to good decisions). As a result, better decisions are more likely to prevail in the next round of the national election. Once the actions have been well-tested they can be institutionalized locally, i.e., they are turned into routine decisions that are once again delegated to distributed controllers. The rest of this chapter will selectively develop this metaphor into a more detailed account.

Before moving on, I want to be clear about the level of analysis I am presently concerned with. The goal of this chapter is to sketch out, qualitatively, the computational theory of the central control mechanism implemented in the BG. Specifically, I will define its goal and function, analyze its component tasks, and spell out the external and internal constraints it operates within. I will focus on the conceptual shape of this computational theory and use empirical literature at the algorithmic and implementational levels as support.

In Section 2, I will argue that the BG function as a central control mechanism because of the role they play in large-scale information integration, flexible decision-making, and robust learning. However, the BG differ significantly from central systems, the archetype of central control mechanisms. This is because, as I will show, the BG utilize a simple message-passing strategy for control (as I will discuss in Section 3), and because the BG do not micro-manage, but train and delegate control decisions to distributed controllers (as I will discuss in Section 4). In section 5, I respond to three objections to my arguments. In section 6, I conclude by discussing how the neurodemocracy account copes with the control problem and its remaining challenges, and by addressing the implications of my account for embodied cognition.

2. Basal Ganglia as a Central Control Mechanism

In this section, I will argue that the BG function as a central control mechanism. I will begin by explicitly articulating and justifying the criteria for a neural mechanism to qualify as a central control mechanism, before offering empirical support for my claim that the BG meet the aforementioned criteria. As we will see, the BG qualify as a central mechanism in a way that is distinctly and interestingly different from the prototypical case of Fodorian central systems.

In considering the criteria for a central mechanism, helpful lessons can be gleaned from Fodor. The Fodorian central systems have the following three features (Fodor, 1983). First, large-scope information integration: central systems are mechanisms where information from all domains can potentially come together (i.e., central systems are, using Fodor's terminology, unencapsulated and domain-general). Second, flexible decision-making: one of the functions of central systems is practical reasoning—they determine which actions have the highest expected utility in a given context, and select them for execution. Finally, robust learning: unlike perceptual or motor modules, which are innate, central systems are where important learning—that is, belief-updating as a consequence of theoretical reasoning—happens.

These three features of central systems form the criteria for central control mechanisms, because together they solve the problems of architecture, coherence, and intelligence discussed in Chapter 1 and Chapter 2. They solve the problem of architecture because unencapsulated and domain-general central systems can mediate between all perceptual and motor modules in order to perform flexible behaviors. They also solve the problem of coherence, because central systems integrate all relevant information to form (largely) unified world and reward models in order to determine the actions to be executed by other modules. Finally, they solve the problem of intelligence by performing theoretical and practical reasoning satisfycingly or even optimally (given Fodor's assumption of ideal non-demonstrative reasoning competence for central systems).

In the following, I will first introduce the basic neuroanatomy and neuroscience of the BG. I will then discuss empirical evidence that suggests the BG meets the aforementioned criteria for central control mechanisms.

2.1. Basal Ganglia 101

The BG are a group of subcortical nuclei with complex connections between them and the other neural structures. These anatomical details of the BG are important for understanding the empirical evidence for my arguments in Section 2, Section 3, and Section 4. As a result, I will need to introduce them briefly before I move on to my main arguments.

The BG (Figure 6.1) are a collection of four subcortical nuclei: First, the striatum is composed of neurons that express D1- and D2-type dopamine receptors (a.k.a. D1 and D2 neurons), among others types of neurons. Second, the globus pallidus contains two parts: an internal part (GPi) and an external part (GPe). Third, the substantia nigra, again, contains two parts, the substantia nigra parts compacta (SNc), and the substantia nigra pars reticulate (SNr). Finally, the last nucleus is the subthalamic nucleus (STN).

Neuronal connections between the BG and the rest of the brain are very complex (Turner & Pasquereau, 2014). For the purposes of this chapter, we will only focus on some of the major connections illustrated in Figure 6.2. First, the striatum and the STN are the major recipients of cortical inputs, which consist of excitatory projections from every region of the cortex (except V1) (Hélie, Ell, & Ashby, 2015, p. 126). Second, major outputs of the BG leave from the GPi/SNr to

the thalamus through inhibitory projections. These inhibitory projections generate action potentials spontaneously at very high rates; as a result, GPi/SNr impose a tonic inhibition on the thalamus. Third, the thalamus also sends excitatory projections back to every region of the cortex. Fourth, the GPe receives inhibitory projections from striatum and excitatory projections from the STN, and projects inhibitory connections to the STN and the GPi/SNr. Finally, the SNc releases dopamine to the striatum, which can have an excitatory effect on D1 neurons and an inhibitory effect on D2 neurons.

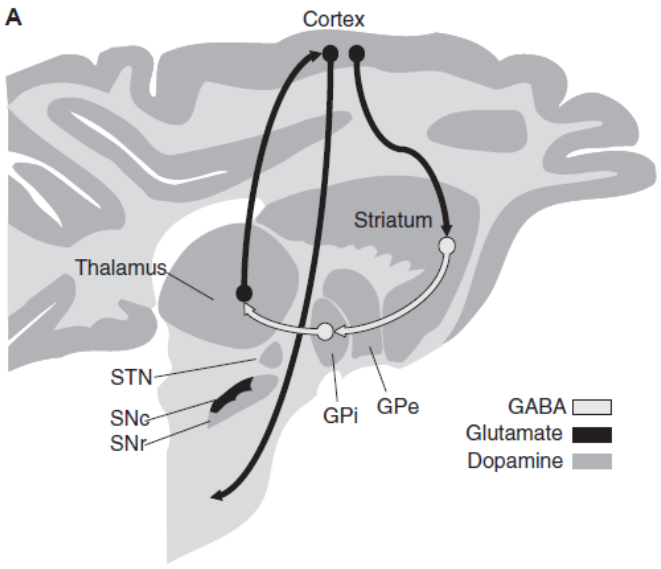

Figure 6.1. Key structures of the BG (in the macaque brain) and some of the major neuronal projections. Excerpted from (Turner & Pasquereau, 2014, p. 436).

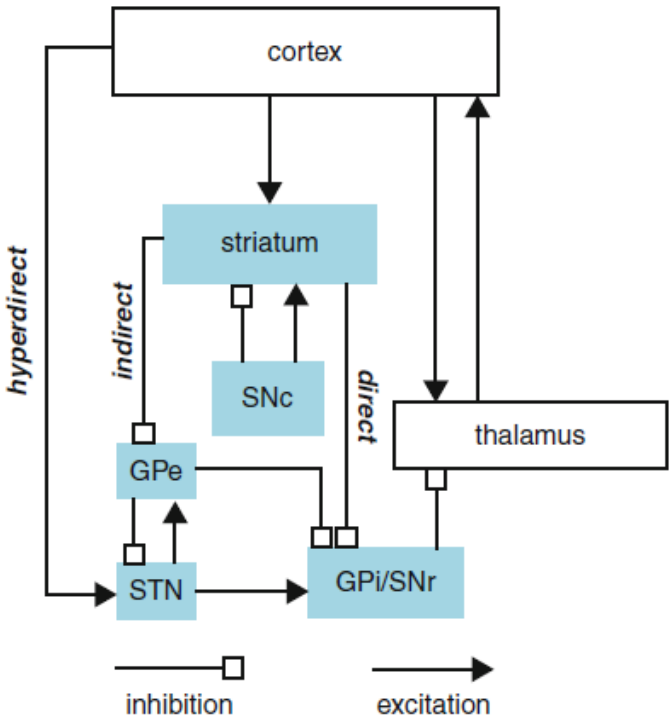

Figure 6.2. Major structures and pathways of the BG. Excerpted from (Jaeger & Jung, 2015, p. 3).

Because the connections between nuclei of the BG are complex, it is most instructive, for our purposes, to consider 3 principle pathways (direct, indirect, and hyperdirect) that run through the BG (Nelson & Kreitzer, 2014; Turner & Pasquereau, 2014). The direct pathway (Figure 6.2) originates from striatum's D1 neurons to GPi/SNr. The net effect of the direct pathway (including one inhibitory connection) is that of inhibiting the BG's output nuclei, GPi/SNr. The indirect pathway originates from the striatum's D2 neurons, through GPe, from which it reaches GPi/SNr either directly or indirectly (through STN). The net effect of the indirect pathway, contrary to the direct pathway, is excitatory for GPi/SNr. The hyperdirect pathway gets its name due to the incredibly fast and excitatory connections from the cortex, through STN, to GPi/SNr. The hyperdirect pathway's reaction time from cortex to the GPi/SNr is shorter than that of the direct pathway. Its net effect, similar to the indirect pathway, is strongly excitatory for GPi/SNr.

Because GPi/SNr exerts a strong and tonic inhibition on the thalamus, activation of the direct pathway, which inhibits GPi/SNr's tonic inhibition, leads to increased thalamo-cortical activity. Increased thalamo-cortical activity, then, promotes the selection of a relevant response option. On the other hand, the activation of indirect or hyperdirect pathways have the opposite effect (i.e., decreased thalamo-cortical activity) and discourage the selection of a relevant response option.

Finally, one unique organizational feature of the BG is the anatomical loop circuits in its pathways. Most prominent is the parallel closed loop circuit: different areas of cortex innervate functionally-specific regions of the BG, which connect to similarly functionally-specific areas in the thalamus, which then projects back to the area of the cortex where the circuit originates (Turner & Pasquereau, 2014) (see Figure 6.3). At a more macroscopic level, these loops are arranged spatially in the BG (and thalamus) in terms of the functional domains they are involved with, such that distant yet functionally related cortical areas send converging projections to the same subregion in the BG (Figure 6.3) (Redgrave, Vautrelle, & Reynolds, 2011).

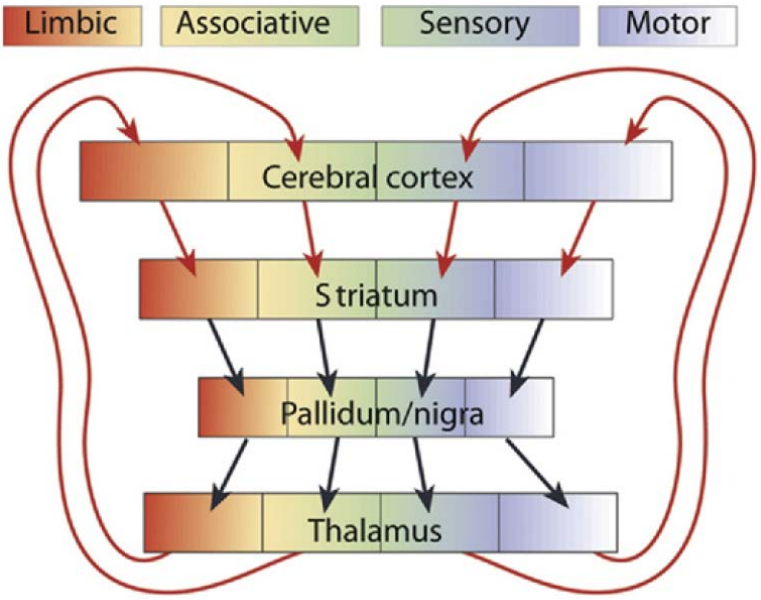

Figure 6.3. Cortical-BG-cortical closed loops in humans. The sensory and the motor domains have been recently revised to be one single sensorimotor domain, and have gained some empirical support from research involving humans and monkeys (Hélie et al., 2015, p. 126; Knowlton, 2015, p. 345). The associative domain is considered to serve the executive function. The limbic domain is considered to serve functions related to emotion and motivation, among others. The list of domains here is not meant to be exhaustive, but rather illustrates the major functional organization of the loop circuits. Excerpted from (Redgrave et al., 2011).

Besides the closed loops, there are also open loops, which provide a pathway for direct communication between functionally distinct areas of cortex (Knowlton, 2015, p. 345). For example, circuits originating from the limbic area can project directly back to the non-limbic frontal cortex.

In the following, I will review the empirical evidence for the BG meeting each of the criteria for central control mechanisms: large-scope information integration, flexible decision-making, and robust learning.

2.2. Large-Scope Information Integration

In this section, I will argue that the BG, similar to the Fodorian central systems, integrate a wide range of information. In fact, the BG can integrate a wider range of information than the Fodorian central systems.

First of all, despite the parallel loop circuits structured into different domains (as discussed above), the internal connections of the BG allow plenty of opportunities for information integration. One obvious opportunity is within the striatum: cortical-striatal projections show some degree of divergence which could allow information from distant cortical areas to be integrated (Hélie et al., 2015, p. 126; Nelson & Kreitzer, 2014).

Another opportunity for information integration is the significant degree of convergence (funneling) of information occurring as the information moves from the cortex, through the BG input (striatum), and toward the BG output (GPi/SNr). The significant degree of funneling is suggested by the anatomical fact that there are far fewer neurons in the BG in general compared to the cortex, and far fewer neurons in BG's output nuclei than the input nucleus (Hélie et al., 2015, p. 126; Turner & Pasquereau, 2014, p. 437).

Moreover, there is an opportunity for information integration in the striatal projection to SNc (dopamine neurons). There is evidence that the limbic loops (Figure 6.3) project from the striatum to all areas of SNc, and the associative loops project from the striatum to all (but limbic) areas of SNc (Knowlton, 2015, p. 346). Additionally, there are more direct empirical observations that the activities of the motor loops are influenced by the activities in the limbic and associative loops (Turner & Desmurget, 2010; Turner & Pasquereau, 2014, pp. 440–443). These two lines of evidence suggest that there is a hierarchy of influence, with the limbic loops having the greatest influence over the others. The associative loops have influence over all but the limbic loops, while the motor loops have only limited ability to influence the other loops (Knowlton, 2015, p. 346).

Finally, recall that Fodorian central systems only receive and integrate output information from perceptual (and proprioceptive) modules and send motor commands to motor (and language) modules in a one-way manner. In contrast, the BG receive and send information reciprocally to both perceptual and motor areas in the brain. The BG also have reciprocal connections with the emotional and associative (executive) areas. As a result, the BG can integrate not only perceptual but also the motor, emotional, and executive (e.g., working memory, attention, planning) information; in addition to this, the BG can directly affect the operation of those areas (Dum, Bostan, & Strick, 2014).

To sum up, empirical evidence suggests that the BG integrate a wide range of information and thus satisfy the first criterion of functioning as a central mechanism. In fact, the BG integrate wider domains of information than Fodorian central systems.

2.3. Flexible Decision-Making

Large-scale information integration serves flexible decision-making—the capacity to select the best options in different contexts reliably—because a wide range of contextual information is essential for making context-sensitive decisions.[112] In this section, I will argue for the empirical case that the BG, similar to the Fodorian central systems, are responsible for flexible decision-making.

There exists a strong consensus and empirical evidence in favor of the idea that the BG function as a motor action-selection mechanism. This owes, in part to the fact that it is easier to conduct experiments correlating motor movements and the BG activities (Humphries, 2015; Prescott, Redgrave, & Gurney, 1999; Redgrave, Prescott, & Gurney, 1999; Turner & Pasquereau, 2014). For example, direct stimulation of the motor regions of the striatum in BG evokes relevant motor movements (Humphries, 2015, p. 352).[113] In addition, there is evidence showing a strong correlation between the BG activities and the onset and parameters of the movements. Finally, the BG's afferent connections from all areas of the cortex allow the action-selection to be sensitive to a wide range of contextual information; the efferent connections, in turn, allow the selected motor action to be implemented in relevant cortical motor areas.

Another important line of evidence for the action-selection function lies in the BG's internal circuitry: the circuitry seems to be designed to implement a selection process (Gurney, Prescott, & Redgrave, 2001a, 2001b). First, evidence suggests that the motor channels (each of which is a subset of the parallel loops belonging to the motor domain) appear to form a somatotopic map (Humphries, 2015). Sub-channels within each channel may also correspond to specific movements of different body parts. This provides a basic anatomical substrate for competition of motor actions. Second, the strong and tonic inhibition connection of the BG's output nuclei (GPi/SNr) to the thalamus effectively block all competing response options and perform the function of "gating." Finally, the direct, indirect, and hyperdirect pathways can work together to select an action option for execution. DVs for different response options from the rest of the brain constitute inputs to loops associated with these options in the three pathways. Then, the three pathways interact in a way to inhibit only the tonic inhibition imposed on the loops that are associated with the selected action (by the BG's output nuclei, GPi/SNr). This results in the inhibitory "gate" opening only for the selected motor action, while remaining a blockade on all other action options. I will discuss in more detail the selection process and its implications in sections 3 and 4.

Moreover, besides "external" motor action-selection, there is strong evidence that the BG select "internal" actions, such as perceptual decision-making (deciding the direction of random dot movement), cognitive action (attention shift and working memory update and maintenance), and the expression of emotion and motivation (Ding & Gold, 2013; Hélie et al., 2015; Nelson & Kreitzer, 2014; O'Reilly & Frank, 2006; Stocco, Lebiere, & Anderson, 2010) (See Figure 6.4) This perhaps is not surprising because the same internal loop circuit structure serving motor selection function is replicated in all other domains, as we discussed in the last section. Moreover, dysfunction of the BG, depending on the specific loops involved, causes disorders at different domains, including disorders in motor movement (e.g., rigidity of movement in Parkinson's

---

[112] This is one of the main reasons Fodorian central systems are not encapsulated: information relevant to a particular non-demonstrative reasoning task can potentially come from anywhere (Fodor, 1983).

[113] Interestingly, lesions of the BG do not seem to impair the motor action completely (Humphries, 2015, p. 352). This fact points to a more complex relation between the BG, the motor cortex, and motor action, which I will discuss in Section 4.

disease), in executive control (e.g., attention-deficit disorder), and in motivation (e.g., obsessive-compulsive disorders), respectively (Nelson & Kreitzer, 2014; Turner & Pasquereau, 2014, p. 441). In sum, the BG are involved in flexible decision-making in all domains, including perception, motor action, cognition, and emotion/motivation.

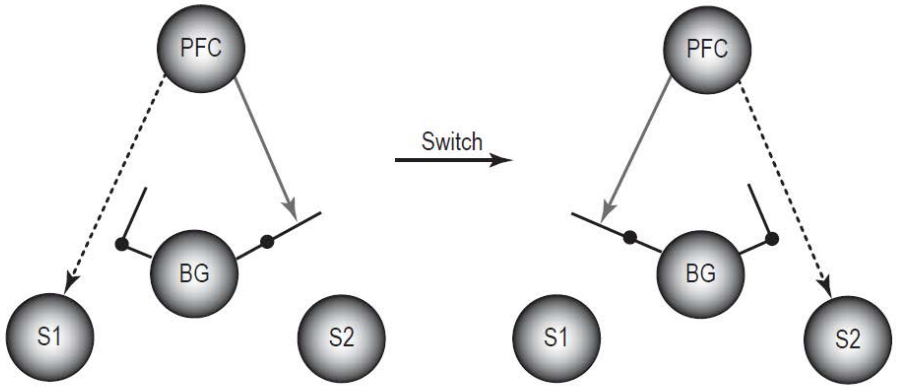

Figure 6.4 Schematic illustration of how the BG select the cognitive action of attention shifting by gating top-down biases from the prefrontal cortex to posterior sensory areas. The PFC biases information processing in the posterior sensory areas with activating or inhibiting signals. The BG select PFC's top-down regulation by gating its influence. Excerpted from (Todd, Hills, & Robbins, 2012, p. 114)

2.4. Robust Learning

To make flexible and context-appropriate decisions, not only do the BG require access to a wide-scope of contextual information, but they also need updated information about the agent and the world as situations change. In this section, I argue that the BG perform the function of robust learning as a central control mechanism because they are the central locus for reinforcement learning (RL), an important type of learning.[114]

Essentially, RL can be seen as a way of biasing response-selection based on reward-related experiences. RL results in behaviors following Thorndike 's law of effect: "any act which in a given situation produces satisfaction becomes associated with that situation so that when the situation recurs the act is more likely than before to recur also" (Thorndike, 1905, p. 203). There are two important ways RL can bias the response-selection.

First, RL can bias response-selection by modifying the value representation associated with a response. The value representation is the predicted estimate of the long-run, expected utility the subject will collect following the performance of this particular response in a particular state (see Figure 6.5). As we discussed in Section 4 of Chapter 5, there are two ways of updating value representations: the first is model-free, and the second model-based. While the BG are clearly implicated in model-free RL, there is evidence suggesting that they are also crucial for model-based RL (Balleine, Dezfouli, Ito, & Doya, 2015; Balleine, Liljeholm, & Ostlund, 2009; McDannald, Lucantonio, Burke, Niv, & Schoenbaum, 2011). However, the detailed mechanisms involved in

[114] For a non-technical treatment of reinforcement learning and its broader implication in philosophy of mind, see (Arpaly & Schroeder, 2014, Chapter 6; Schroeder, 2004)

both model-free and model-based learning are currently unknown (Redgrave, Gurney, Stafford, Thirkettle, & Lewis, 2013).

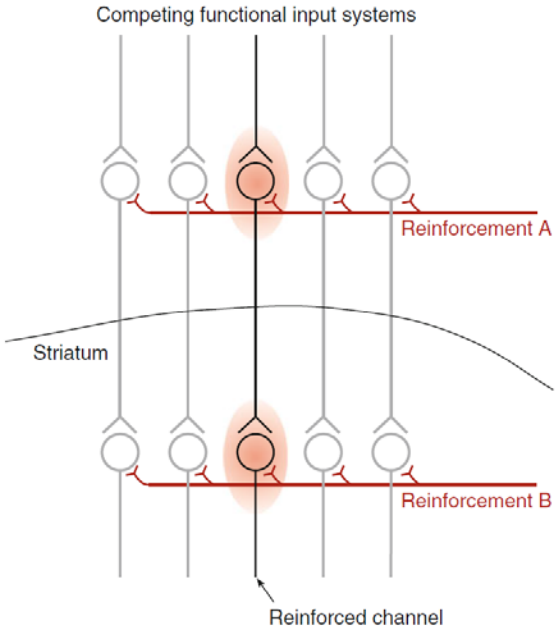

Figure 6.5 Two ways RL can bias the response-selection. Reinforcement A works by updating the value representation for response options (through either model-based or model-free mechanisms) before they form inputs to the BG. Reinforcement B works by changing the striatum's sensitivity to different information processes. Each neuron in the figure represents a population of neurons (or a neural mechanism). Excerpted from (Redgrave et al., 2013)

Second, RL can bias the response-selection by modifying the sensitivity to the inputs from different information processes (Figure 6.5). We know much more about the neural mechanisms subserving this type of RL. The BG is the central locus for this type of RL. Crucial information for the RL, the reward prediction error, is carried by the temporary (phasic) change in the amount of dopamine released by SNc to the striatum (Knowlton, 2015). Specifically, the effect of phasic dopamine change on the striatum's D1 neurons (involved in the direct pathway) is the inverse of the effect it has on D2 neurons (involved in the indirect pathway): On the one hand, the phasic increase in dopamine associated with an action (indicating more than expected rewards) strengthens the cirtico-striatum (D1) connections in the associated loops. As a result, it increases D1 neurons' sensitivity to the relevant information processes that have promoted the selection of this action. On the other hand, a phasic increase in dopamine weakens the cortico-striatum (D2) connections in associated loops. As a result, it decreases D2 neurons' sensitivity to the relevant information processes that have discouraged the selection of this action.

Because the sensitivity to information processes promoting the selection of this action is increased and the sensitivity to information processes discouraging the selection of this action is decreased, the net effect of a phasic increase in dopamine is that the BG will be more likely to select this action again next time under the same context. On the contrary, the phasic *decrease* in dopamine (associated with less than expected reward) has a net effect that BG will be less likely to select the relevant action next time under the same context. In short, the second type of RL improves action-selection by changing the BG's sensitivity to relevant information processes in direct and indirect pathways such that an action option becomes more likely to be selected if it leads to more rewards than expected, and vice versa.

As discussed earlier, the action-selection in the BG involves not only external motor actions, but various internal actions as well. So, RL can potentially adjust not only the cognitive system's disposition to perform different motor actions, but also various dispositions involved in cognition (e.g., the disposition to pay attention to different stimuli, and to recall different information from long-term memory), perception (e.g., the disposition to making perceptual judgments about the world), emotion/motivation (e.g., the disposition to elicit different emotional responses). In fact, there is evidence suggesting the BG are involved in the RL of motor and cognitive responses of different abstractions in the brain (i.e., moving one's leg this way or that way, turning left or right, and foraging or resting) (Badre & Frank, 2011; Frank & Badre, 2011; Hélie et al., 2015; Ito & Doya, 2011). In sum, the BG are crucial for various types of RL, which can adapt the cognitive system's internal and external actions flexibly to the agent's/organism's changing environment.

In this section, I have shown that the BG meet all three of the criteria associated with central controllers: large-scope information integration, flexible decision-making, and robust learning. Nevertheless, the BG differ from the Fodorian central systems significantly. One obvious difference is that there are two BGs in the brain and, as a result, there are two central controllers with identical functions. In the final section of this chapter, I will briefly discuss an implication this has for the control problem. The BG markedly differ from the central systems in another way: they are involved not only in tasks traditionally construed as decision-making—such as practical reasoning and theoretical deliberation—but also in decision-making involved in motor, perceptual, and motivational tasks. Moreover, unlike a central system, the BG do not micro-manage all decisions; on the contrary, they act as a kind of 'tutor' for the cortex. While the BG are involved in decision-making in novel contexts, they slowly train the cortical and the other subcortical mechanisms, and eventually transfer the relevant decision-making to them as the decisions become habitual. I will discuss this "tutor" account of the BG in more detail in Section 4. In the next section, I first wish to highlight the most significant difference between the BG and central systems: the BG are simple message-passing controllers.

3. The Basal Ganglia as a Simple Message-Passing Controller

In this section, I will argue that the BG utilize a simple message-passing strategy, rather than a rich message-passing strategy, to coordinate neural mechanisms. I will first briefly discuss the differences between these two strategies. Then, I will show that the BG constitute a simple message-passing controller because they rely on simple internal mechanisms and simple signals of activation/inhibition for their control function.

3.1. Rich Message-Passing and Simple Message-Passing Strategies

As we've touched on in Chapter 1, the distinction between rich and simple rich message-passing strategies is coined by Clark (1998, 2001). On the one hand, control mechanisms utilizing a rich message-passing strategy use representations that have rich or detailed content to coordinate other mental mechanisms. These control representations are usually amodal or in general-purpose format. In addition, they also tend to rely on rich internal models of the world, as well as complex construction and transformation of representations, for their control function. For example, the central systems in Fodorian architecture coordinate between perceptual and motor modules using amodal representations with detailed perceptual or motor content. Central systems also process these internal representations with complex algorithms (if one does not accept Fodor's conclusion that central systems cannot be naturalistically explained). We should note that the use of a rich message-passing strategy does not imply the use of centralized control strategy. For example, the massively modular architecture discussed in Chapter 3 utilizes a rich message-passing strategy with the distributed control strategy.

On the other hand, control mechanisms utilizing a simple message-passing strategy, such as the hierarchical neural control structures in HECA, use control representations whose content is simple (such as activation or inhibition). Such control mechanisms tend to rely on specialized internal models, which may operate on highly proprietary or special-purpose formats. This is because when neural mechanisms communicate with simple messages only, there is less pressure for them to share the representational format of their internal models. They also need not rely on complex information processing for their control function. For instance, in the subsumption architecture, the higher-level mechanism coordinates the lower-level ones through signals that only function to encourage or discourage their activities (Brooks, 1991a, 1991b).

Although Clark does not emphasize this, we should note that the distinction between the two strategies should be drawn at both the algorithmic and the implementational levels (Marr, 1982). At the algorithmic level, one specifies the representations for the input and output of a particular mechanism, as well as the algorithms for their transformation. At the implementational level, one specifies how these representations and their transformation can be realized physically. The rich-message vs. simple-message distinction concerns the control representation's content (i.e., at the algorithmic level) as well as the control representation's realizer (i.e., at the implementational level).

As a result, one cannot argue that this distinction is not meaningful on the basis that, given a neural implementation, all basic units interact with each other through inhibition or activation. The fact that a neural implementation consists of units that inhibit and activate one other does not imply that the *content* of the implemented representation is activation or inhibition. Let me illustrate this point with two examples. First, a representation with rich content can be implemented by a connectionist or neural model consisting of neuron-like units that interact, for the most part, by inhibiting and activating one another's activities. The same representation can also, however, be implemented by classical computational architecture that lacks these features of mutual activation and inhibition (say, a Turing machine). Second, a representation carrying the simple content of activation or inhibition can be implemented with either a classical computational architecture or a neural network. Additionally, there are multiple ways to implement the representations of inhibition and activation within a neural network: for example, by deploying a different combination of inhibitory and excitatory connections between neurons. It is a special case and an example of an efficient design that the BG implement the inhibition signal (at the algorithmic level) with a strong tonic neural inhibition (at the implementational level).

Having clarified the distinction between a rich message-passing and simple message-passing controller, I will now provide empirical support for my claim that the BG is a simple message-passing controller.

3.2. Basal Ganglia: A Simple Message-Passing Controller

Most computational and neuroscientific models support a simple message-passing account of the BG (Gurney et al., 2001a, 2001b; Hélie et al., 2015; Humphries, 2015; O'Reilly & Frank, 2006; Stewart, Bekolay, & Eliasmith, 2012; Turner & Pasquereau, 2014). In Section 5.3 I will address the only model I know of that considers BG a rich message-passing controller (Stocco et al., 2010) and argue that it is not empirically supported.

In the following, I will discuss one of the most well-known and representative computational models of the BG (Gurney et al., 2001a, 2001b), and illustrate how the BG work as a simple

message-passing controller.[115] As we will see, the BG rely on highly simple and specialized internal mechanisms and utilize simple control signals for their control function. I will begin by illustrating how the BG perform the control function with internal mechanisms that rely on specialized internal models and simple information processing.

According to Gurney's model of the BG, the channels in various domains, which constitute the parallel loops running through the BG, form the basis for response-selection. Roughly speaking, we can view each channel as corresponding to each response option. For example, one channel in the motor domain may correspond to the option of moving one's arm upward, while a neighboring channel corresponds to the option of moving one's arm leftward, etc. So, the selection of one response option over competing alternatives is done through the selection of a particular channel.

Two qualifications are needed before we move on to the process of selection: first, each channel in the BG may not correspond to just *one* internal or external response option, but *many*. That is, the selection of a channel may correspond to the selection of multiple response options at the same time. Second, I avoid claiming that each channel "represents" an action; it is more accurate to say, rather, that each channel "corresponds to" or "indexes" an action (Hardcastle & Hardcastle, 2015). This is because these channels do not carry detailed information about the actions they index. They do not, for example, carry the visual information to be updated into working memory, or the motor action plan to be executed, etc.). Instead, the channels only carry minimal information in the sense that, roughly speaking, if a channel in the BG is selected, the corresponding response option in its corresponding neural mechanism will be executed.

To return to the selection process: the three major pathways—direct, indirect, and hyper-direct ones—interact to select the best option by creating a "winner-lose-all" dynamic. First, decision variables (DVs) for each response option are computed by different information processes subserved by cortical and other subcortical mechanisms and sent through excitatory connections to the striatum's D1 neural population (the direct pathway), the striatum's D2 neural population (the indirect pathway), and STN (the indirect pathway) respectively (see Figure 6.6). In Gurney's model, the identical DVs are sent to the three different pathways.

Second, the interaction between the direct and hyperdirect pathway creates an off-center, on-surround network that exhibits a "winner-lose-all" dynamic in BG's output (GPi/SNr) nuclei. As illustrated in Figure 6.7, through the direct pathway, the striatum's D1 neural population in each channel projects a focused inhibitory input to an SNr/GPi population within the same channel. In contrast, through the hyperdirect pathway, the STN population in each channel projects a more diffuse excitatory input to SNr/GPi, influencing neural populations in the same but also the neighboring channels. The net effect of the interactions of the direct and hyperdirect pathways is that the channel receiving the strongest cortical input, due to the strong and focused inhibitory output from the striatum's D1 neural population and relatively weaker excitatory output from STN, ends up with the lowest firing rate in its corresponding SNr/GPi population (hence, "winner-lose-all"). On the contrary, all of its neighboring channels, due to the relatively strong counteracting excitatory output from STN, and weaker inhibitory output from the striatum's D1 neural population, will have much higher firing rates.

[115] In the next chapter, I will discuss another well-known model that differ importantly from Gurney's (O'Reilly & Frank, 2006; Wiecki & Frank, 2013). However, both of them support the simple message-passing controller interpretation of the BG.

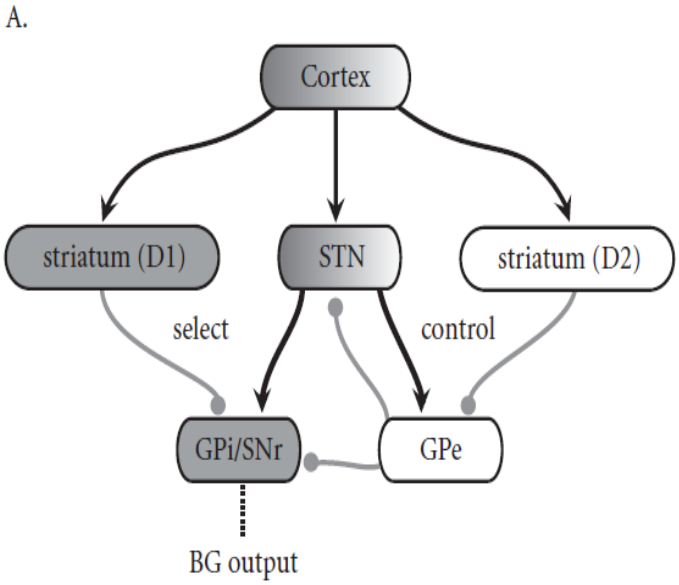

Figure 6.6 Gurney's (2001a, 2001b) BG model. The black arrows indicate excitatory connections; the gray arrows indicate inhibitory connections. Excerpted from (Eliasmith, 2013).

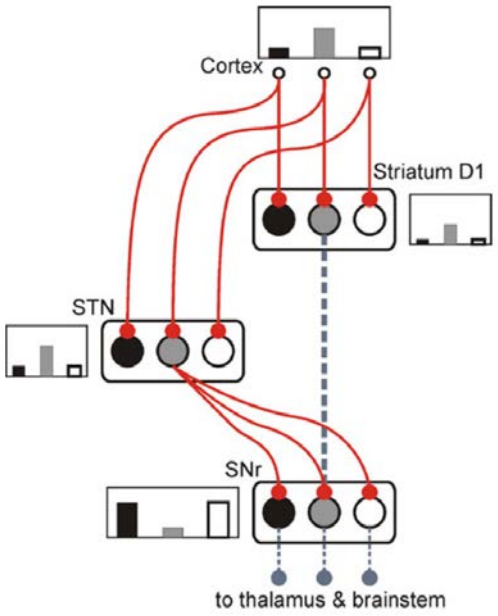

Figure 6.7 The off-center, on-surround network at SNr/GPi. The bars in each box represent the neural activities of different neural populations in each neural structure. The solid lines represent excitatory connections; the dashed lines represent inhibitory connections. Inhibitory connections from two of the striatum's D1 populations to the corresponding SNr/GPi populations within the same channels are omitted. Diffuse excitatory connections from two of the STN populations to the SNr/GPi are omitted as well. Excerpted from (Humphries, 2015).

Let me illustrate in more detail how the off-center, on-surround network is realized through a more simplified, non-dynamic, model (see Figure 6.8). In this simplified model, cortical mechanisms provide DVs of 0.3, 0.8, and 0.5 respectively to populations corresponding to three competing options in STN and the striatum's D1 neural populations (all through excitatory

connections with a connection weight of 1). From the striatum's D1 neural populations to GPi/SNr, the signals travel through focused inhibitory connections with a weight of -1 respectively (the negative weight reflects the inhibitory connection). From STN to GPi/SNr, the signals travel through diffuse excitatory connections with a weight of 0.5 each. The channel that receives the highest DV of 0.8 from the cortical mechanisms ends up with the lowest activity (i.e. 0) in its GPi/SNr population, while the other channels maintain a positive value in their GPi/SNr populations.[116]

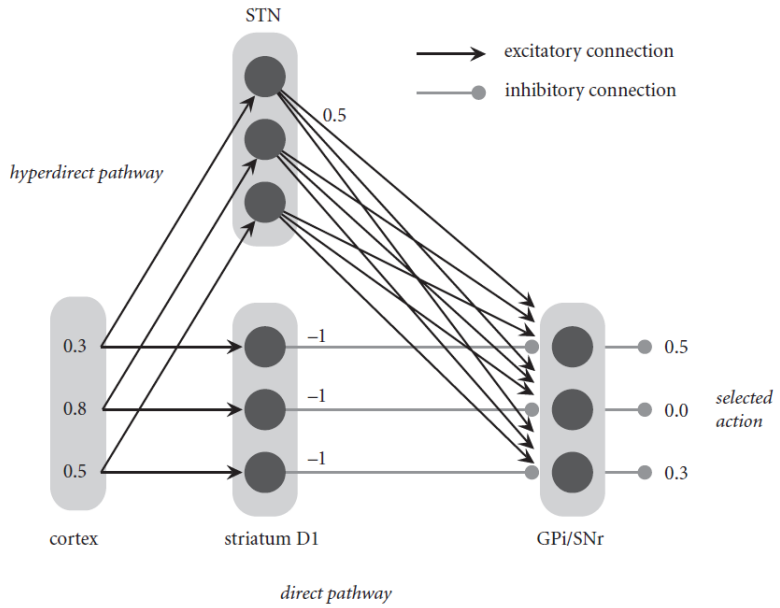

Figure 6.8 A simplified BG network.

Note that the off-center, on-surround network with a "winner-lose-all" dynamic is achieved through a fine balance of strong, focused, inhibitory connections from the striatum's D1 neural population, with weak, diffuse, excitatory connections from the STN. Without the diffuse excitation from the STN, all populations in the GPi/SNr will be equally suppressed (with a value of 0, because neural populations in GPi/SNr cannot represent negative values). However, this simplified network only works well with carefully picked DVs and a restricted number of options. If the DVs from cortical mechanisms for all options are all similarly high or low, or if the numbers of options are large, this simplified model will not exhibit the desirable off-center, on-surround feature. As a result, Gurney et al. (2001a, 2001b) have argued that the indirect pathway is necessary for a modulating "control" function,[117] in addition to the "select" function performed by the direct and hyperdirect pathways (see Figure 6.6): the indirect pathway enables the GPi/SNr to exhibit the desirable "winner-lose-all" dynamics across a wide range of different situations.[118]

---

[116] The GPi/SNr value for the channel receiving the highest DVs from the cortex is calculated in the following way: 0.8*-1 (from Striatum) + 0.5*(0.3+0.8+0.5) (from STN) = 0.

[117] The "control" function in this model is not to be associated with the control problem discussed throughout this thesis.

[118] For a simple explanation of how the indirect pathway performs this "capacity scaling" function, see (Eliasmith, 2013, p. 166; Humphries, 2015, p. 351).

It is worth noting, before we move on, the similarities between this BG model and the non-linear leaking competing accumulator (LCA) model introduced in Chapter 5—which, recall, is primarily a model for action-selection mechanisms in the cortex (Usher & McClelland, 2001).[119] Because this BG model can be seen as a value-based, multi-hypothesis generalization of the non-linear LCA model (which implements the sequential probability ratio test (SPRT) introduced in Chapter 4), the BG's action-selection mechanism can be seen as implementing a rational model of Bayesian inference that trades off speed and accuracy optimally (Bogacz, 2007).[120] Similarly, this also explains why intelligent decisions can emerge from the dynamical and epistemically cooperative processes in the BG's action-selection mechanisms, which accumulate DVs from a large scope of heterogeneous information processes (as discussed in the Cooperative Decision Thesis Section of Chapter 4).

In short, unlike the cortex, which is "an information processing resource that is dedicated to manipulating, remembering, binding, and so forth representations of states of the world and body" (Eliasmith, 2013, p. 170), the BG's internal mechanisms are much simpler—the BG neither specify nor evaluate the response options. That is, they select the best responses from response options specified by other action-specifying processes, through a "winner-lose-all" dynamics, with simple DVs provided by other action-evaluating processes. This is in sharp contrast to the Fodorian central systems, which specify and evaluate response options with carefully constructed internal world/reward models and complex reasoning competence.[121]

Having illustrated how the BG perform the control function, I will now demonstrate that they utilize a simple message-passing strategy for control. As already discussed, the BG select the best response option through a "winner-lose-all" dynamic: only the channel receiving the highest DVs will have its associated SNr/GPi neural activities paused temporarily, while neural populations of all other competing channels will maintain their positive neural activities. Because SNr/GPi's inhibitory connection to the thalamus has a very high tonic (spontaneous) firing rate, the temporary pause leads to a temporary disinhibition of the thalamic neural population in the same channel. As illustrated in Figure 6.2, this allows relevant information in that thalamic neural population (which

---

[119] This BG model is a value-based, multi-alternative generalization of the Non-Linear LCA model, which only accounts for two competing options for perceptual decision-making (i.e., a non-value-based decision that assume a fixed utility for making the correct perceptual discrimination). This BG model's hyperdirect pathway (with its diffuse projection to all competing channels) serves the function of providing a background countering activation, which is realized by the reciprocal recurrent inhibitory connections between competing accumulators in the Non-Linear LAC. (I am ignoring the complication that the non-linear LCA is a dynamical recurrent model, while the BG model I discuss here is a simpler, feed-forward model.) Also, the indirect pathway's modulation makes the BG model perform better than the Non-Linear LCA model across a wider range of situations, e.g., in novel or complex contexts where there are likely to be more options and DVs for different options are likely to be similar. Gurney et al. argue that this is one of the reason why the BG play an important role in novel or complex decision-making (2001a, 2001b). I will discuss this in more detail in the next section.

[120] For a discussion on the cortico-BG-thalamus-cortical loops as implementing a Bayesian inference procedure called the Empirical Bayes, see (Eliasmith, 2013, p. 281).

[121] Valerie Gray Hardcastle and Kiah Hardcastle (Hardcastle & Hardcastle, 2015) argue that our current understanding of the BG selection mechanism challenges Marr' algorithmic level of analysis, because the BG mechanism does not transform a representation via some algorithm, but merely "chooses which representation to enact." However, I do not find their argument sound. Even if the BG mechanism does not construct and transform complex representations, it is clear that BG at minimum perform some simple transformation of simple representations: the value representations (DVs) at each channel are transformed (accumulated) and it is on the basis of the transformed value representations that certain channels are selected over the others. As a result, the BG mechanism does not pose a threat to Marr's algorithmic level of analysis.

originates from a cortical or other subcortical areas) to be relayed back to its origin, and leads to the selection of the corresponding response.[122] That is, the SNr/GPi communicates with the thalamus with signals of inhibition and disinhibition (at the algorithmic level of analysis), which is in fact implemented by the strong tonic neural inhibition and temporary disinhibition of the thalamus (at the implementational level of analysis).

The means by which the BG exercise control with simple commands can be illustrated using a simplified example. In Figure 6.4, the BG mechanism controls the information flow from the prefrontal cortex through simple commands. The simple commands work as gating signals that (through the process discussed above) either allow or prevent information (which can be either rich or simple) from the prefrontal cortex to travel to other brain regions. This is, again, in sharp contrast with central systems, which control other modules by issuing detailed representations.

To conclude this section, unlike the Fodorian central systems, the BG are a centralized, simple message-passing controller. The BG utilize control representations with simple contents (inhibition and disinhibition) to coordinate the other neural mechanisms in the brain. Additionally, the BG rely on relatively simple internal mechanisms for the control function: instead of building up detailed, general-purpose models of the world and reward contingency, BG merely weight and accumulate DVs for different options, relying on other mechanisms for action-specification and action-evaluation. By integrating evaluative information from a wide range of sources to inform the selection of a wide range of actions, the BG enhance the coherence and intelligence of the cognitive system.[123]

4. The Interactions Between the Basal Ganglia and the Distributed Controllers

Another important difference between the BG central controllers and the Fodorian central systems is this: unlike the central systems, which micro-manage and never delegate important decisions, the BG central controllers do not select every action, but only those at the early stages of skill acquisition. Subsequent to this, they act as "tutors" and train the distributed controllers to perform the same actions in those contexts. In the following, I will begin by elaborating this "tutor" account of control and its empirical evidence, after which I will explain why the BG are uniquely qualified for the role of tutor. I end with a discussion of the novel implications this account has for how we think about the neural substrates of flexible and inflexible/habitual behaviors and decision-making.

According to the tutor account of the BG (Ashby, Ennis, & Spiering, 2007; Ashby, Turner, & Horvitz, 2010; Hélie et al., 2015; Turner & Desmurget, 2010; Turner & Pasquereau, 2014, p. 445), the BG are not involved in every action-selection (internal or external), but only those at the early stage of skill acquisition and learning. That is to say, the BG are only involved *before* a particular action or action sequence becomes habitual and can be elicited automatically by certain stimuli in a way that lacks flexibility and context-sensitivity. In fact, there is a division of labor between the BG and the other distributed controllers. The BG central controllers learn to select the

---

[122] See (Humphries, 2015, p. 352) for a detailed example of how the disinhibition process work.

[123] The BG model we discuss in this chapter remains quite simplistic: It assumes the three pathways receive the same signals from the cortex. In addition, it does not model the interactions between channels of different domains that do not compete with each other directly. These interactions may be required for the BG to play a more competent role in control by integrating large-scope information in more complex ways. In the next chapter, I will consider a different model that focuses on these interactions in slightly more detail. However, I do not think the additional interactions within BG we know of, or those we may discover in the future, will affect the simple message-passing interpretation of the BG mechanism significantly.

optimal and context-appropriate action option in new contexts. Once the actions-selection are well-practiced, other parts of the brain (e.g., the cortex and other subcortical areas) will take over the storage, initiation, and production of these now habitual behavior patterns.[124]

Empirical support for the tutor account includes emerging evidence suggesting that the BG are necessary for learning and making novel and flexible response-selection, while habitual and reflexive response-selection requires only the involvement of distributed controllers (Turner & Desmurget, 2010; Turner & Pasquereau, 2014, pp. 444–445). For example, empirical studies show that while patients with BG lesions often have problems learning new skills and sequences of actions, interruption of the BG activity (through injection of an inhibitory chemical into the BG's output nuclei GPi/SNr) does not significantly affect a well-practiced, habitual motor response: for example, the motor response's initiation, response time, direction, and real-time error-correction are unaffected, and its velocity and the extent of the movement is only mildly affected. On the other hand, inactivation of the relevant cortical motor region induces significant impairments of habitual motor response.[125]

This division of labor can potentially be explained by the different types of action-selection and learning mechanisms involved in the BG and the cortex: the BG's more complex action-selection

---

[124] It is interesting to note that one group (Turner and his colleagues) that supports the tutor account of the BG also argues against the thesis that the BG is a central action-selection mechanism. One of the main reasons the BG cannot be viewed as a central action-selection mechanism is, they argue, that discharge in the GPi/SNr (output of BG) related to the disinhibition of a particular motor action (i.e., the BG's selection of the motor action) begins later than the earliest relevant muscle activation and the activation of primary motor cortex. As a result, the BG cannot be the mechanism responsible for the selection of the particular action, but can only play a modulating role (Turner & Desmurget, 2010; Turner & Pasquereau, 2014). However, this piece of evidence will work against the BG as central action-selection mechanism only if human cognitive systems follow the classical sandwich model, where the central mechanism selects an action prior to the detailed specification of the motor program (which often involve activities in both the primary motor cortex and relevant effectors). As discussed in Chapter 4 and Chapter 5, we have good empirical and theoretical reasons to think that classical sandwich model does not capture how human cognitive systems are organized. In fact, action-evaluation and action-specification happen simultaneously. As a result, GPi/SNr's slightly later discharge (which is still earlier than the actual movement) is no evidence against BG as a central mechanism responsible for action-selection.

[125] The tutor account has become increasingly dominant in recent years. Yet, there are several other existing competing account, some of which are compatible with the division of labor between of the BG and distributed controllers advocated in the tutor account (Balleine, Dezfouli, Ito, & Doya, 2015; Liljeholm & O'Doherty, 2012). There is one prominent account (Yin & Knowlton, 2006) that is not compatible with this division of labor. It claims that the development of automatic behavior involves the transfer of control from associative striatum to sensorimotor striatum (instead of from the BG to other distributed controllers). According to this account, the associative striatum (Figure 6.3) is responsible for the learning of novel skills, and the sensorimotor striatum is responsible for controlling the automatic, habitual skills. Supporting evidence for this account includes empirical data suggesting a double dissociation: temporary inactivation of the sensorimotor striatum does not impair the learning of new motor sequence but impairs the execution of previously acquired motor sequence; meanwhile, temporary inactivation of the associative striatum impairs learning of new motor sequences but does not impair the execution of acquired motor sequence as much. This alternative account, however, is not compatible with the empirical evidence we discussed earlier that temporary inactivation of the BG impairs little the execution of the well-practiced, automatic behavior. There are many ways to square the evidence supporting this alternative hypothesis with our tutor account (Ashby, Turner, & Horvitz, 2010; Hélie, Ell, & Ashby, 2015); however, I will not be able to go into this as it will go beyond the scope of this chapter. It is worth noting that the core claim that BG are a central mechanism will still hold even if we assume this alternative account is true, and the tutor account is false. It is because this alternative account attributes the BG the control of habitual behaviors, in addition to the learning and decision-making of novel behaviors. In other words, it attributes more control functions to the BG (thus, making the BG an even more powerful central control mechanism) than our tutor account suggests.

mechanism, as well as their fast, feedback-based RL capacity, make them a suitable tutor for the distributed controllers.

First, Gurney et al. (2001a, 2001b) argue that the BG's action-selection mechanism is necessary for novel decision-making. Decision-making in the initial stage of skills acquisition often involves a large number of response options, many of which have similar DVs. (This is perhaps because the cognitive system has not learned through experience which response option is the best and which can be ignored). As discussed in the last section, the BG's internal circuitry, with the indirect pathway playing the "control" function, is capable of reliable and speedy selection of the response option with the highest values (DVs) in situations involving a large number of options with similar values. However, the relatively simpler cortical action-selection mechanism is incapable of such reliable and speedy selection in these situations (Gurney et al., 2001a, 2001b). On the other hand, habitual decisions and inflexible behaviors can depend only on cortical action-selection mechanisms. This is because habitual behaviors rely on solidified knowledge of a relatively narrow-scope of stimulus–response mapping that has been fine-tuned (by the BG, as we will discuss next) and stored in the cortex. This remains the case even if the mapping is quite complex and requires the rich and hierarchical information processing of the cortical areas.

Second, BG's fast learning ability, in combination with their context-sensitive decision-making capacity, enable them to discover the most optimal and context-sensitive response options quickly and train the distributed controllers. While the BG utilize RL, the cortical mechanisms utilize the Hebbian learning mechanism, which is relatively slow and not sensitive to the feedback signals (Hélie et al., 2015). According to the Hebbian learning rule, the strengthening of a given action in a situation, or a stimulus–response association, depends on repetition rather than its consequences, e.g., whether it leads to positive or negative results. Specifically, a Hebbian learning process will strengthen synapses with strongly correlated pre- and post-synaptic activities and weaken synapses with weakly correlated activities. In short, for Hebbian learning, neurons that fire together wire together.[126] Roughly, we can think of the pre-synaptic activities as representing a stimulus, the post-synaptic activities as representing a response, and the synapse between them as implementing the relevant stimulus–response mapping. If a stimulus (partially) causes a response, Hebbian learning, regardless of its reward consequences, will make it even more likely that the stimulus leads to the response next time.

This is why the BG are important for tutoring the cortical mechanism. Fast RL and context-sensitive decision-making enable BG to learn to select the optimal response more quickly and more reliably through trial-and-error with reward-related feedback. When the BG select the optimal response option, they also activate the corresponding post-synaptic target (representing the response) in the cortex given the preceding pre-synaptic signals. Therefore, by selecting the context-appropriate action over and over again, the BG ensure that the appropriate cortico–cortical synapses will be strengthened through Hebbian learning, while the inappropriate ones will be weakened (we can think of the BG as helping tune the distributed controllers). Once the right cortico–cortical connections are firmly established, the BG are no longer required to produce the action and will leave the storage and execution of the stimulus–response mapping to the distributed controllers.

---

[126] More accurately: "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth or metabolic change takes place… such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 1952, p. 50). Note that the pre-synaptic activities need to (partially) cause the post-synaptic activities within the right time window for Hebbian learning to happen.

Empirical evidence from studies on working memory, rule-guided behavior, cognitive and motor skills learning supports this account (Ashby et al., 2007, 2010; Hélie et al., 2015; Turner & Desmurget, 2010). Computational modeling explicitly based on this account of the interaction between the BG and the cortex also fits empirical data well (DeWolf & Eliasmith, 2013). In short, the BG are responsible for tutoring the distributed controllers by helping establish the appropriate stimulus–response mapping, and transferring the more flexible and context-sensitive BG decision-making to the more habitual and inflexible decision-making in the cortical and other subcortical areas.

Finally, the tutor account of BG has an interesting implication. It suggests that the line between flexible and inflexible decision-making is surprisingly drawn between flexible decision-making that involves both the BG and other brain areas on the one hand, and inflexible decision-making that does not involve the BG on the other.

Typically, the line between flexible and inflexible decision-making is drawn, as it is in the Standard Account of classical cognitive science, between central systems and modules, It is worth noting that Fodor's account of cognitive architecture does not allow for the distinction, introduced by the dual-process theory of reasoning (Evans, 2008), between type-1 (automatic) and type-2 (deliberate). This is because all judgments and reasoning processes in Fodor's account involve central systems. However, it is probably fair to suggest, as some theorists have, (e.g. Samuels, 2005), that the relevant division can be drawn between information processes involving either only perceptual and motor modules or separate reactive modules (the automatic processes), and processes involving perceptual modules, central systems, and motor modules (the deliberate processes).

As a result, the tutor account of BG gives us a picture of cortical and subcortical information processes that is very different from how we ordinarily think about them: we typically think that the evolutionarily new cortical areas (especially the prefrontal cortex) are associated with sophisticated information processes and flexible behaviors, while the evolutionarily old subcortical areas (including the BG) are associated with simple information process and stereotypical behaviors. On the contrary, the BG literature suggests that flexible and context-sensitive decision-making, which depends on large-scale information integration and sensitivity (including sensitivity to reward prediction error for performing the RL), requires the involvement of the BG and a wide range of cortical and subcortical areas.

To recapitulate, in the last three sections I have argued, first, that the BG's main function is to solve the control problem by implementing a central, simple message-passing controller. In fact, compared to the Fodorian central system, which only integrates perceptual or quasi-perceptual information, the BG integrate a larger scope of information. Moreover, the BG make decisions in more domains than Fodorian central systems, which do not make decisions in the sensory, motor, or emotion domains. Finally, the BG enable learning in more domains than traditional central systems, which only learn in the central cognitive domain (that is, in belief-updating). Importantly, the BG, instead of issuing rich commands, exercise control by using simple commands that perform a gating function. However, the BG specialize in the selection of novel, context-appropriate actions, as well as the entrainment and subsequent delegation of habitual behaviors to the distributed controllers. In short, I have offered a hybrid account of how central and distributed controllers, using a simple message-passing strategy, work together in the cognitive system. In addition to this, I have provided some empirical support for the claim that human cognition in fact depends on central controllers, although not those postulated by classical cognitive science.

5. Objections

In this section, I address three objections to the thesis that BG implement a central control mechanism utilizing a simple message-passing strategy. The first objection, often advanced by more theoretically-inclined embodied cognitive scientists, is that cognitive systems do not need central controllers to generate coherent and intelligent behaviors. The second objection, coming from a more empirical point of view, is that the BG controllers are indistinguishable from (at least some) distributed controllers; as a result, they do not deserve to be called "central" controllers. The third objection, made by more classically-oriented cognitive scientists, is that the BG implement a rich message-passing controller. As we will see, addressing these objections will help clarify and demonstrate the theoretical and empirical strengths of my account.

5.1. Coherent Behaviors Can Result from Distributed Control

It could be argued that the BG do not constitute a central mechanism because, in general, cognitive systems do not need central mechanisms to produce coherent behavior. This type of objection is prominent in the embodied cognitive science literature, perhaps because the notion of a central control mechanism is reminiscent of the problematic "Cartesian intellect." In the following, I will examine some of the arguments against the necessity of central control mechanisms for producing coherent behaviors. I will show that these arguments have no force against my core claim that the human cognitive system depends on the BG central controllers for generating novel coherent behaviors. Finally, I will show that we can reconcile the embodied cognitive scientists' urge to argue against the central control mechanisms and the empirical facts they use to do so with the fact that the BG is indeed a central controller.

Clark (1998, 2001) has argued that coherent behaviors can result from the self-organization of cognitive mechanisms and the influence of distributed control structures that are internal or external to the cognitive systems. These influences include the global dissipative effects of neuromodulators, external constraints from social, institutional, and technological scaffoldings, and the biasing influence of neural control structures that utilize a simple message-passing strategy for control. Importantly, these control structures contribute to the coherence of the cognitive system using what Clark calls "ecological control," which "does not micro-manage every detail, but rather encourages substantial devolvement of power and responsibility" (Clark, 2007, p. 2). That is, these distributed control structures do not dictate the goals to the controlled cognitive mechanisms. In contrast, the controlled cognitive mechanisms remain more-or-less autonomous in the sense that they have some say about their own goals (I've discussed this conception of autonomy in Chapter 4). The distributed control structures "nudge and tweak" cognitive mechanisms in certain directions by activating or inhibiting their activities. In short, central mechanisms are not necessary for coherent behaviors, because:

> [The coherent cognitive system] is nothing but the cumulative effect of the co-active unfolding of the various resources supporting different aspects of adaptive response. This unfolding is determined by a delicate mix of sparse ecological control and pure self-organization. (Clark, 2007, p. p21)

There is some evidence that supports the claim that coherent behaviors can emerge from distributed control. First, there are existence proofs that coherent behaviors can emerge without a central mechanism. Empirical studies of cnidarian nervous systems show that many forms of coherent behavior can be achieved without centralized neural structures (Prescott, 2007). Also, it seems that "the ant colony has no boss, and no virtual boss either, and gets along swimmingly with distributed control…" (Dennett, 2007, p. 96).

Moreover, computational cognitive scientists and roboticists have successfully modeled competent action-selection and decision-making with distributed control (Prescott et al., 1999; Redgrave et al., 1999; Thagard, 2001). In particular, the *recurrent reciprocal inhibition* (RRI) is a principle exploited by these models. RRI architecture (Figure 6.9) is observed in many different areas in both vertebrate and invertebrate brains. It can be seen as a multi-hypotheses generalization of the non-linear LCA model discussed in Chapter 4 and Chapter 5. The architecture consists of competing nodes (populations of neurons representing competing options), each of which has inhibitory connections to every other (hence, the recurrent reciprocal inhibition). These nodes also have excitatory links from their inputs, as well as excitatory links to the effectors. This architecture exhibits an effect of positive feedback: increased activity in a node will increase its inhibitions on all other nodes, which in turn decrease their inhibition on it. The net effect is a "winner-take-all" functionality—the conflict among competing options is resolved when the node that has some minor advantage in its activity end up inhibiting all other nodes and take control of the effector entirely. The "winner-take-all" characteristic of RRI makes it a good architecture for implementing action-selection. To model a more complex pattern of action-selection, the input connections to nodes and the reciprocal inhibitory connections can be fine-tuned, and additional non-competing nodes and excitatory connections between them can be added. In short, the success of this model in performing coherent action-selection suggests that coherent behaviors can emerge from cognitive systems lacking a central mechanism.

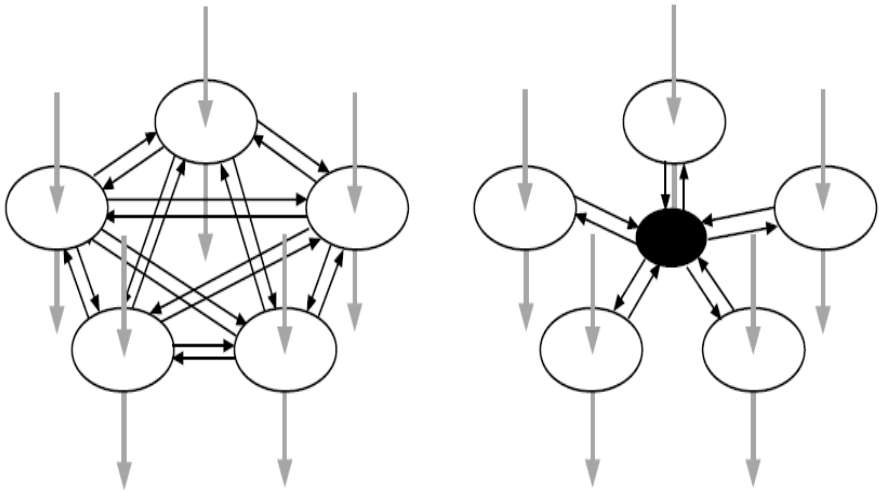

Figure 6.9 Distributed recurrent reciprocal inhibition architecture (left) and central selection architecture (right). Dark arrows represent inhibitory connections, and light arrows represent input and output excitatory connections. Excerpted from (Redgrave et al., 1999).

Even if we grant the truth of all the empirical literature mentioned above, this objection is invalid. First, all the evidence reviewed here amounts to an existence proof that coherent behaviors *can* result from a system with distributed control. This evidence may constitute a good objection against the claim that central mechanisms are *necessary* to produce *all* coherent behaviors in *all* cognitive systems. However, I have made no such claim. All I have argued for is the much more modest conclusion that *human* cognitive systems (and perhaps those of other vertebrates) *in fact* take advantage of the BG as a central mechanism to produce *some* coherent behaviors. As the tutor account of the BG suggests, the BG may not be involved once behaviors become habitual; as a result, some habitual coherent behaviors in human cognitive systems may not depend on the BG.

Second, all the evidence discussed above concerns either simpler cognitive systems (e.g., cnidarian nervous systems) or sub-systems within a complex one— for example, the BG as a sub-system also

utilizes a more complex action-selection mechanism similar to RRI (Prescott et al., 1999). In contrast, many large-scale cognitive architectures aiming to model human cognition, such as "ACT-R" (Adaptive Control of Thought—Rational) (J. R. Anderson, 2007) and "Spaun" (Semantic Pointer Architecture Unified Network) (Eliasmith, 2013; Eliasmith et al., 2012), build in the BG as a central selection mechanism. These facts suggest that perhaps only simpler cognitive systems or sub-systems can produce coherent activities or select actions without central mechanisms—a claim that I do not have to, nor want to, dispute.

In fact, we can reconcile the fact that simpler cognitive systems can behave coherently without a central control mechanism with the fact that complex cognitive systems require a central control mechanism for coherent behaviors. Gurney and his colleagues (Prescott, 2007; Prescott et al., 1999; Redgrave et al., 1999) suggest two related reasons why central mechanisms are preferred over purely distributed control in a large-scale and complex cognitive system. First, complex cognitive architectures with central control mechanisms are more economical to build and easier to increment as they develop. For example, RRI architecture requires n(n-1) connections to arbitrate between n competitors and requires additional 2n connections to add a new competitor (in the most extreme scenario).[127] On the other hand, a system with a central mechanism requires only 2n connections to arbitrate between n competitors, and requires only additional 2 connections to add a new competitor (see Figure 6.9). As a result, the connection-cost benefit for a system with a central mechanism quickly outweighs the initial cost of building a central mechanism as the system gets larger. This is because the connection cost for a system with a central mechanism is lower than that for RRI architecture when n is larger than four; also, the total connections required for RRI architecture increase much faster than the total connections required for a system with a central mechanism. In short, one of the design principles behind the central controllers may be that they provide a significant advantage for the economy of connections and for the ease of increment for larger and complex cognitive systems. Because simpler cognitive mechanisms may not benefit as much from central control mechanisms, this may explain why they do not have one.

Moreover, we can also reconcile embodied cognitive science's anti-Cartesian mistrust of central mechanisms with my claim that the BG is a central controller. Anti-Cartesian cognitive scientists worry about positing central mechanisms because they tend to be reminiscent of the Cartesian intellect, a "central controller" that acts as the locus of consciousness, self-control, knowledge, and/or rationality. However, my account is much more minimal and does not consider the BG central controller to be the sole implementing mechanism of any of these capacities, although the BG central controller is very likely to contribute to all of them (Barron & Klein, 2016; Haggard, 2008; Merker, 2007).

To conclude, the fact that some simpler cognitive systems can produce coherent behaviors without central mechanisms does not constitute evidence against my claim that the BG is a central mechanism that is in fact involved in (and potentially necessary for) generating coherent novel behaviors in human cognitive systems.

5.2. Basal Ganglia Controllers Are Not Different from Distributed Controllers

The second objection to my account is that the BG controller is not a *central* controller because there is nothing special about it—it is functionally similar to distributed controllers. Against this, I

[127] There are ways to reduce the amount of connections necessary, e.g., through a small-world network architecture (Shanahan, 2012).

will show that there is indeed something unique about the BG. My critics fail to see this uniqueness because they've lumped together two importantly different sub-tasks in decision-making: action-evaluating and action-gating. Once this distinction is spelled out, it is clear that the BG constitute a *global action-gating* controller, while distributed controllers are (at best) *global action-evaluating* controllers. I will end by explaining away a related intuition that the central controller, if there is one, should be implemented by the evolutionarily-new PFC, instead of the evolutionarily-ancient subcortical BG.

This second objection can be stated briefly as a *reductio ad absurdum* argument:

(1) Suppose that the BG controllers are "central" controllers.

(2) The BG controllers are no different from (at least some) distributed controllers.

For example, distributed controllers in the PFC engage in large-scale information-integration, learning (even some types of reinforcement learning as well), and decision-making, just like the BG.

(3) Hence, if one were to argue that the BG controllers are central controllers, one would have to call many distributed controllers central controllers too. The distinction between central and distributed controllers then become meaningless.

This is a *reductio ad absurdum*.

(C) As a result, the BG controllers are not central controllers.

Let me unpack this argument by turning to an empirical theory that makes an argument of the same form. According to the *affordance competition hypothesis* (Cisek, 2012a, 2012b; Cisek & Kalaska, 2010; Cisek & Pastor-Bernier, 2014), the BG are not qualitatively different from the other neural control structures. Rather, they all perform what Cisek considers the action-selection function in a decision-making task. As I have discussed the general framework of the affordance competition hypothesis in the last chapter, in what follows I will focus only on the relevant details.

First of all, Cisek distinguishes two sub-tasks in a decision-making task: action-specification and action-selection. (It is important to note that Cisek's terminology of action-selection refers to a subtask of a decision-making task, while I have used action-selection as a synonym for decision-making. I will adopt Cisek's terminology here temporarily.) Action-specification concerns the specification of (some) potential options and their implementational details. For example, in a decision concerning arm movements, the direction and speed of different potential arm movements need to specified. Action-selection concerns the selection of a potential option for execution based on the DVs generated by action-evaluating information processes or mechanisms.

Second, Cisek makes explicit claims about the sub-task of action-selection: the sub-task depends on the winner-take-all dynamics discussed earlier, as well as biases (e.g., DVs) from distributed evaluative mechanisms in the brain. For example, a motor action may receive DVs from the orbitalfrontal cortex for the objective economic value of the outcome of a particular action option and the lateral PFC for its conformation with context-dependent rules (Cisek, 2012b, p. 930; Cisek & Pastor-Bernier, 2014, pp. 9–10). An action option is selected when its neural activities become strong enough to inhibit those of all competing options.

As a result, according to this model, BG and other evaluative mechanisms in the PFC are all involved in the same sub-task of action-selection in decision-making, providing inhibitory or

excitatory signals that biases the competition of action options. As Michael L. Anderson, following Cisek's account, maintains:

> … rather than conceive of basal ganglia as a gatekeeper in a central, specialized action-selection circuit… it is arguably better to think of basal ganglia as one important source of biasing inputs that can influence ongoing pattern competition between different response opportunities. (M. L. Anderson, 2014, p. 223)

In short, because BG controllers perform the same action-selection function as distributed controllers, they cannot be uniquely distinguished as central controllers.

In response to this objection I shall argue that BG are indeed unique: although they perform the action-selection function, BG are not action-evaluating mechanisms. Cisek and M. L. Anderson fail to see this because their task-analysis is inadequate: the action-selection sub-task should be analyzed into two functionally distinct operations of action-evaluation (biasing) and action-gating. Specifically, BG are unique because they are a "global gating controller" while other neural controllers (such as those in the PFC) are at best "global biasing controllers." BG's global gating function, together with their large-scale information integration and robust learning, is what make them a central controller. In the following, let me unpack this in a more detailed manner.

An adequate task-analysis of action-selection should decompose it into at least the following two subtasks: action-evaluation and action-gating (Kable & Glimcher, 2009; Rangel, Camerer, & Montague, 2008).[128] Action-evaluation involves the process of assessing available action options for their values (by approximating, for example, their expected subjective utilities) and generating DVs in order to influence the decision-making. Action-gating, on the other hand, involves the process of comparing different options and selecting the one with the highest DVs. For example, the inputs to the RRI architecture discussed earlier can be seen as DVs generated by various evaluative mechanisms to bias the decision-making. The winner-take-all dynamics of RRI architecture, in particular the positive feedbacks generated by reciprocal inhibitions among competition options discussed above, serves a gating function to select the option with the highest DVs.

The BG are global *gating* controllers because they are responsible for gating actions in many different domains. Distributed controllers are only responsible for gating the actions they specify. For example, distributed controllers in the motor cortex (or for Cisek, the sensorimotor loops) specify relevant motor response options and perform a gating function to select one option for execution. Other distributed controllers will specify different actions (e.g., those in the dorsolateral PFC may specify options for more abstract goals), but they also only gate the actions they specify (Cisek, 2012b; Shadlen & Kiani, 2013, p. 797). BG, however, are capable of gating actions in many different domains, before these actions become habitual.

In contrast, distributed controllers in the PFC are global *biasing* controllers. Because they are located at the higher-level in the feedforward/feedback neural hierarchy (as we discussed in Chapter 4), they can bias the decision-making of all the lower-level mechanisms directly or indirectly. However, the PFC controllers, like other distributed controllers, are not global gating controllers: they can only gate their own cognitive actions (of action-evaluation or biasing).

How do we explain the intuition, held by many neuroscientists and philosophers, that if any neural controllers should be considered *central* controllers, those implemented in the PFC are the best

[128] The terms used here are different from those used by the original authors, but they refer to the same sub-tasks.

contenders? I think this intuition is rooted in research that identifies mechanisms in the PFC as the neural correlates of the "executive function." Executive function includes various higher-cognitive capacities, such as planning/deliberation, the inhibitory and flexible control of behaviors, and working memory (Diamond, 2013). Empirical research has identified the neural correlates of these capacities in various areas of the PFC. The ability to represent contextual information and abstract goals to facilitate context-appropriate actions has been found to correlate with activity in the lateral PFC; the ability to evaluate the values of objects with activity in the orbitofrontal cortex; and finally, the ability to plan future actions flexibly and think counterfactually with activity in the ventromedial PFC and dorsomedial PFC (D'Esposito & Postle, 2015; Dixon & Christoff, 2014; Koechlin, 2016).

I do not deny that neural mechanisms in the PFC have greatly enhanced the intelligence of human decision-making. However, they contribute to producing more intelligent behaviors as global biasing controllers, by providing DVs to other distributed controllers and the BG central controllers. As Patricia Churchland and Chris Suhler put it:

> By embellishing the ancient subcortical reward system organization [i.e., the BG] with fancy cortical input, a plan can be evaluated for its likely consequences. Richer cortical input allows for richer predictions and evaluations. Goals can be nested within goals. Plans can become very elaborate and goals very abstract. (Churchland & Suhler, 2014, p. 318)

In addition, there is evidence that the BG are heavily involved in these executive function, such as working memory and response inhibition (Balleine et al., 2015; D'Esposito & Postle, 2015; O'Reilly & Frank, 2006). This should not be surprising, because we already know that BG are involved in training all cortical controllers in novel situations. Moreover, some of the unique challenges that make executive function necessary include the overriding of habitual behaviors in order to perform flexible, context-sensitive behaviors in novel situations. It is no wonder that executive function relies heavily on the BG as well the PFC.

In conclusion, the BG controllers are indeed unique among neural controllers. They *gate* actions globally, while the other controllers (at most) *evaluate/bias* actions globally. BG's global gating capacity, together with the large information-integration and robust learning capacities, is what make the BG controllers deserve the name of central controllers.

5.3. Basal Ganglia Are a Rich Message-Passing Controller

In the last two sections, I dealt with objections against the claim that the BG are central controllers. In this section, I address an objection to my claim that the BG are *simple message-passing* controllers. In doing so, I will discuss the only account that models BG as a rich message-passing controller (J. R. Anderson, 2007; Stocco et al., 2010) and argue that, given the empirical literature, this model is unlikely to be true.

This account of BG, according to J.R. Anderson and his colleagues, is inspired by the procedural module (i.e., the production system) of the ACT-R cognitive architecture (J. R. Anderson, 2007). According to ACT-R, the procedural module constantly monitors and recognizes the information in the cortex, and selects appropriate actions to perform (i.e., by sending appropriate information as inputs to other modules). The procedural module functions as if it is following production rules that can be expressed by a symbolic, IF-THEN statement. Both the antecedent and the consequent of the rule can be quite complex, so the rules need to be expressed by propositions connected by logical operators and implemented in symbolic format (at least at the algorithmic level). In short,

the BG are considered to be the neural structures that subserve the procedural modules (J. R. Anderson, 2007).

In a recent computational model, John R. Anderson and his colleagues (Stocco et al., 2010) show how the BG controller can perform the function of a procedural module as illustrated in the following example. First, the striatum of the BG can recognize information from the cortex in order to select which IF-THEN rule is to be followed. Second, the consequent of the rule is constituted by a routing operation that can be expressed as "route the representation X from source S to destination D." Third, there are cortico-cortical pathways that connect many cortical areas (e.g., from S to D). However, the strength of the signals (e.g., representation X) travelling from source S through the cortico-cortical pathways are usually not strong enough to affect the operation at destination D.[129] Finally, the BG can transmit an additional (compressed) copy of representation X through the BG-thalamic-cortical pathway to destination D.[130] This redundant, compressed copy of representation X from the BG will provide the necessary additional strength needed for destination D to process the representation X from source S. That is, it essentially allows BG to gate the operation at the destination D.[131]

Anderson *et al.*'s model is a rich message-passing model of BG because it utilizes a rich message (the compressed copy of representation X) for gating and coordinating cognitive mechanisms. It also relies on relatively rich models and the complex transformation of amodal representations for its control function. For example, the striatum needs to rely on models that can map complex patterns of cortical states to the appropriate routing action.

I will only offer a brief reply to this objection because I do not think it has much empirical merit. Anderson *et al.*'s model, as I mentioned earlier, is the only model that I know of that offers a rich message-passing interpretation of the BG. However, this model is largely motivated by an ACT-R framework and not backed up by empirical evidence. For example, the authors acknowledge that this model is developed using ACT-R's procedural module "as a reference for the functional properties to implement in the circuit" (Stocco et al., 2010, p. 552). Furthermore, they concede that the model "takes an original stance on the role of the indirect pathway" (Stocco et al., 2010, p. 552)—specifically, they take its function to be that of determining the destination of the conditional routing, which is an essential part of a rich message-passing operation. In adopting this stance, however, the authors cite no supporting empirical literature.

In addition, the rich control signals posited by this model (i.e., the compressed copy of representation X passed from BG to the cortex via the thalamus) is incompatible with a known empirical fact about the SNr/GPi's connection to the thalamus. As we discussed earlier, the SNr/GPi projects a tonic and highly inhibitory connection to the thalamus. This specific neural connection is unlikely to implement a communication channel that transmits signals richer than inhibition/activation because its strong tonic inhibition is likely to shut down completely whatever information processing that happens in the corresponding thalamic regions. As a result, this neural

---

[129] ACT-R also have subsymbolic operations, such as numerical values that "control strength and accessibility of those symbolic structures" (Byrne, 2012, p. 435).

[130] The authors recognize that the numbers of neurons in BG are significantly lower than those in cortex. As a result, BG cannot represent cortical information in its original format. Whatever representations from the cortex need to be significantly compressed before reaching the BG (Stocco, Lebiere, & Anderson, 2010).

[131] (Stocco et al., 2010) provides a simple example of how the routing process works.

connection is good for implementing a simple message communication channel of inhibition and disinhibition, but is not good for anything else.

Finally, even if Anderson *et al.*'s model of the BG were true (which is highly unlikely given the evidence), it would not support the typical classical cognitive science approach of centralized control with a rich message-passing strategy. This is because this model falls short of portraying the BG as utilizing the prototypical rich message-passing strategy. Specifically, this model incorporates a key feature from the simple message-passing strategy: the rich control signals (e.g., the compressed copy of the representation X) work by *activating* the relevant operations in the cortical mechanisms, rather than by transferring the input representations (e.g., motor intentions) to be processed directly. In short, the rich message-passing model of BG is currently empirically implausible and theoretically *ad hoc*.

To sum up this section, I've shown that we have good empirical ground to believe that the BG are a centralized, simple message-passing controller. The BG controllers are distinct from distributed controllers because they work as global gating controllers. As such, they play an important role in orchestrating intelligent and coherent behaviors in complex cognitive systems.

6. Conclusion

In this final section, I will return to the metaphor of the BG as the central election commission in a neurodemocratic society of mind, in order to provide a qualitative summary of the BG's computational theory. Then, I will discuss in more detail the progress this account of neurodemocracy makes in addressing the control problem, focusing on the contribution of the BG central controllers. Finally, I end by discussing the broad implications to the embodied mind approach.

In a neurodemocratic society of mind, habitual decisions are determined through local elections that are held by local governments of distributed controllers. Yet, when novel and complex decisions arise that cannot be settled locally, the BG central controllers will intervene. The BG controllers differ from the folk-psychological conception of a neural executive-in-chief that is enshrined by Fodorian models. Unlike Fodorian central systems, the BG controllers do not have beliefs, desires, and rationality; nor do they specify candidate actions or evaluate them. Instead, the BG are more similar to a central electoral commission that holds national elections. The commission is impartial to all candidates, because it does not evaluate nor represent them in detail. Yet, it has the necessary infrastructure to run big elections: registering multiple proposed candidate actions, collecting "votes" from all walks of life, and reliably determining which candidate actions have the highest votes (which corresponds to the large-scope information integration and context-sensitive decision-making). That is, the BG facilitate the national elections so that they result in more intelligent and coherent final decisions. More precisely how the BG central controllers do this will be addressed shortly.

Once the winning candidate actions are determined, the central election commission does not implement the action itself, but simply allows the local governments to implement the chosen actions (which corresponds to the BG's simple message-passing strategy). However, the BG central controller does enforce accountability: it tracks the results and hold voters accountable for their votes—voters whose votes contribute to worse than expected actions will have their votes counted less the next round, and vice versa. This process will improve the decision-making through rounds of elections until a relatively good decision is found. (This corresponds to the RL function of the BG.)

Finally, once the right actions are found, local institutions can be set up to make these decisions routinely. The BG help tune the appropriate weights given to "voters" participating in the local election so that local decisions can be made quickly and reliably (which corresponds to the role of BG as a tutor of other distributed controllers). In short, the BG function as a central controller in a bizarre yet well-designed way.

Now it is time to revisit the control problem and evaluate how my neurodemocracy account helps resolve the problems of coherence and intelligence.[132]

The BG address the problem of coherence in several ways. Most importantly, the BG subject novel response-selection across levels in various neural hierarchies to the direct influence of evaluative information from many domains. This has two benefits for coherence. The first is that the BG central controllers can orchestrate response-selection at various levels in the hierarchies directly and simultaneously. Specifically, the BG can orchestrate response-selection more efficiently than the highest level of distributed controllers, whose influence needs to be mediated by neural mechanisms at their lower-level and take time to travel down the hierarchies. Second, all decisions made in the BG are likely to involve large and overlapping sets of evaluative mechanisms; responses selected on the basis of such consensus are potentially more likely to cohere with each other. It is intuitively plausible, for example, that a set of decisions are more likely to be mutually coherent if we subject them to the majority vote of a single group than if we subject each decision to the majority vote of different groups—assuming, that is, that the different groups have different goals and beliefs.

Moreover, the BG controllers have a more sophisticated selection mechanism compares to the distributed controllers (partly due to the complex interaction of the three pathways as we discussed earlier). As a result, the BG can inhibit unselected responses completely while disinhibiting the selected responses fully. This capacity prevents incoherent behaviors due to two or more conflicting responses being both selected (disinhibited) at the same time (Gurney et al., 2001a, 2001b).

Finally, because responses that conflict with others tend to lead to less rewarding outcomes, RL can also work to eliminate the selection of such responses. Importantly, the BG can be seen as relying on a reiterative process that improves the coherence of behaviors through cycles of response-selection and RL. Relatedly, the BG also indirectly contribute to the coherence of behaviors by training distributed controllers to select habitual responses, whose neural, bodily, and environmental consequences are conducive to overall coherent behaviors. For example, developing habits to work in a less distracting environment or learning to maintain focus on one thing at a time.

We should note that these contributions to coherent behaviors remain true even if the brain has two BG. That is, because the two BG involve large and overlapping sets of evaluative mechanisms, they will tend to select responses that are coherent with each other, compared to distributed controllers

[132] The BG controllers also help address the problem of architecture. As discussed in Chapter 1, the problem of architecture concerns how the neural connections among the cognitive system's highly-distributed neural mechanisms should be set up to enable flexible coalition-formation and problem-solving. This problem can be addressed by the connective core architecture, a set of densely-connected hub nodes that are topologically central in the network of the cognitive system, which allows highly distributed neural mechanisms to communicate with each other (Shanahan, 2012). The BG are a part of the connective core (due to the topologically central location and wide connection); specifically, the BG, with their specialized action-selection mechanism capable of subserving multiple competition simultaneously, serve as a "locus of competition" for forming neural coalitions.

at the higher-level which integrate less information.[133] In addition, the other features—RL, sophisticated selection mechanisms, and direct influence over mechanisms at different hierarchies—remain conducive to coherence regardless of how many BG there are.[134]

BG also help resolve the problem of intelligence. As I discussed earlier, the BG (approximately) implement a multi-hypothesis version of the sequential probability ratio test (SPRT). As a result, BG's decision-making process can increase the reliability of decision-making by accumulating and comparing DVs for different options. Moreover, the BG also help overcome the challenges of managing dynamics, capacity, and accuracy. As discussed in the final section of the last chapter, these challenges need to be overcome for the BG's response-selection mechanism to select the optimal options reliably. Moreover, one of the common strategies for overcoming these challenges is assessing the accuracy of DVs to inform relevant control processes. Through RL, the BG can enhance their assessment of the accuracy of DVs generated by different information processes. For example, the BG can address the challenge of managing accuracy and increase the reliability of response-selection by weighing DVs according to their expected reliability.

However, it is clear that RL is not sufficient for solving the problem of intelligence entirely. This is because RL can only improve the assessment of the accuracy of DVs in contexts with which the subject has extensive experience. Additional control processes need to be implemented by the BG or other neural structures in order to achieve more reliable response-selection in novel contexts. In the next chapter, I will suggest possible ways in which the BG might implement such control processes in order to enhance the intelligence of decision-making.

I would like to conclude this chapter by briefly addressing an apparent tension between my account of neurodemocracy and embodied cognition. My argument that the BG function as a central controller may seem at odds with the embodied mind approach, which often advocates the complete decentralization of control. This is because at the core of the embodied mind framework is a vision of cognition as a bag of tricks—in Clark's words, cognition is "a mixed bag of relatively special-purpose encodings and stratagems" (Clark, 2001, p. 100). My account of the BG, however, is compatible with this vision. What I have shown is that the central controller does not need to be fashioned after Fodorian central systems; it need not be the ultimate source of intelligence, nor must it rely on rich representational models for its control function. Instead, the central controller is just one additional "trick"—one that utilizes a simple model and specializes in bringing together the intelligence of other good tricks distributed across the neural system in order to tackle complex and novel problems. In the next chapter, I will suggest a different framework for thinking about the BG, and consider a few more "tricks" they might implement in order to better cope with the distinctive challenges of novel contexts.

---

[133] There are contralateral cortico-striatal projections, although the ipsilateral ones are more predominant (Goldman, Compte, Wang, & Squire, 2009, p. 97). Also, the two BG do seem to have direct neural connections with each other, at least in rats (Bak, Markham, & Morgan, 1981), in addition to indirect ones where they communicate with each other indirectly through other neural mechanisms.

[134] In fact, the absolute numbers of the central controllers may not be the key determinant of coherent behaviors. If there are 10 BG controllers, all of which are equipped with these relevant features for generating coherent behaviors discussed above, they will still enhance the coherence of behaviors despite the large numbers. However, to equip a cognitive system with 10 BG controllers with the kind of wide-spread neural connections and internal mechanisms would be too expensive biologically.

7

Conclusions and Further Research

1. Conclusions

The goal of this thesis is to understand how human cognitive systems solve the control problem—the problem, that is, of how distributed and diverse neural mechanisms coordinate with each other in the production of intelligent and coherent behaviors. In Chapter 1, I analyzed the control problem into three sub-problems. The problem of architecture concerns the infrastructure for control. To solve it, neural connections between neural mechanisms must somehow be set up appropriately and efficiently so that they can be flexibly recruited to form "neural coalitions" to tackle a wide range of existing or novel problems. In addition to concerns about "control infrastructure," there are concerns about "control decisions." The problem of coherence refers to the challenge of making coherent control decisions in the numerous and distributed neural mechanisms in order to produce large-scale and goal-directed behavior. Similarly, the problem of intelligence concerns the issue of how intelligent control decisions can emerge from the interaction and coordination of less intelligent neural mechanisms.

After clarifying the control problem, I considered its solution space. I showed that it is a two-dimensional space with four conceptually distinct positions. Besides the traditionally well-known solution from classical cognitive science (centralized control with a rich message-passing strategy) and the recently popular solution of embodied cognitive science (distributed control with a simple message-passing strategy), there are two lesser known alternatives. The first, employed by the massively modular architecture, combines distributed control with a rich message-passing strategy, while the other marries centralized control with a simple message-passing strategy. In the rest of the thesis, I evaluated the strength and weakness of each of these four positions and developed my own neurodemocracy account—a hybrid solution that pairs centralized and distributed control with a simple message-passing strategy.

Part I of my thesis focused on arguments against the massive modularity (MM) architecture. I showed that the MM architecture cannot overcome an important type of control problem, the information control problem. As a result, it cannot produce human-level intelligence. In order to provide a better theoretical context for understanding the MM hypothesis, I also introduced the Standard Account, which is typically associated with the classical cognitive science.

I began Chapter 2 by introducing the intuitive conception of behavioral flexibility as a key feature of human intelligence. According to the intuitive conception, humans have the competence to perform tasks satisficingly in a context-appropriate fashion, and in a wide range of novel and significantly different contexts. I then reviewed the Standard Account's epistemic commitment (ideal non-demonstrative reasoning competence) as well as its architectural commitment (central systems). Despite the Standard Account's merits—its potential to overcome the problems of architecture, coherence, and intelligence, as well as to explain human intelligence—it faces fatal problems on both empirical and theoretical grounds. What we learned from the Standard Account's mistakes is that we need to retreat from this extremely idealized model of cognition and human reasoning that assumes optimal performance and does not consider internal and external constraints on biological cognition. It was against this background that we were best placed to appreciate MM. The rest of Chapter 2 introduced MM as a paradigm that models human cognition more realistically and with more careful consideration of resource and time constraints. Architecturally speaking, MM is committed to the idea that the human mind is composed exclusively of a massive set of Darwinian modules, which are domain-specific and informationally-encapsulated computational mechanisms. More importantly, MM's epistemic commitment is the heuristic approach to human reasoning. According to the heuristic approach, human cognition utilizes heuristics, principles or algorithms for information processing that provide satisfying outcomes within the constraints of tractable computation.

In Chapter 3, I introduced the confederate account of MM, which incorporates learning and self-organization, in order to explain how modules can (learn to) self-assemble to generate behavioral flexibility. Having introduced the confederate account, I showed that it is beleaguered by an explanatory gap. In order for the confederate account to general behavioral flexibility, information needs to be routed to the correct module at each step of the information processing. However, it is not obvious that MM architecture possesses the control competence necessary for the relevant information control in a wide range of significantly different novel contexts. I argued that MM cannot fill this explanatory gap because of its nativist commitment. For an extreme version of MM incompatible with learning, the innate control competence cannot handle information control in a wide range of novel contexts. While the moderate version is compatible with learning, its learning capacity needs to be significantly constrained by innate information, which would prevent the acquisition of the control competence necessary for behavioral flexibility. In other words, MM cannot overcome the information control problem because it cannot deal with the problem of intelligence. Given MM's nativist commitment, the interactions of less intelligence modules cannot generate the intelligent control that is necessary for behavioral flexibility. While Part I constituted a predominantly negative project, the positive take-home message was that in order to understand flexible control mechanism, we need a mechanistic account of self-organized control that is anti-nativist and implements heuristics.

In Part II of the thesis, I constructed the hierarchical embodied cooperative architecture (HECA)—an empirically-updated society of mind account. Then, I evaluated it as an embodied cognitive science solution to the control problem, which utilizes distributed control with a simple message-passing strategy. In Chapter 4, I began by reviewing Dennett's Pandemonium architecture. I showed that, insofar as it posits both distributed control and a simple message-passing strategy, the Pandemonium architecture is consistent with the approach of embodied cognitive science. The

upshot of this, however, is that its solution to the control problem shares the same weaknesses: it is unclear that the problem of coherence and the problem of intelligence can be solved by the competitive interactions of distributed neural agents.

Subsequent to my critique of the Pandemonium architecture, I presented HECA and clarified its three theses. First, the Embodied Agent Thesis develops Dennett's neural agents into action-specifying and action-evaluating embodied information processes. Second, the Hierarchical Structure Thesis contends that neural mechanisms are structured hierarchically. They represent the causal structures of the world at different depths, and/or process information with computations of varying complexity. As a result, neural mechanisms subserve different levels of information processes that vary systematically in their flexibility, capacity, and speed. Finally, according to the Cooperative Decision Thesis, intelligent decisions emerge from the dynamical and epistemically cooperative process of accumulating DVs from heterogeneous information processes until a decision threshold is met. I concluded the chapter by placing HECA on the center left of a spectrum of cognitive architectures which ranges from radical embodied cognition (the far left) to the Standard Account (the far right). HECA is a moderate version of the embodied cognitive architecture because it supports embodied cognition's key contention that cognitive capacity constitutively depends on the sensorimotor region of the brain. HECA is also compatible with embodied cognition's more ambitious idea that cognitive capacity constitutively depends on the body and the environment. However, it is not compatible with embodied cognitive science's most radical idea of anti-representationalism, because representations—especially modality-specific and action-centric ones—play important roles in HECA.

In Chapter 5, I reviewed the empirical literature in order to defend HECA and furnish it with more mechanistic details. The empirical literature I reviewed included: the affordance competition hypothesis; hierarchical models of perception and action-control; the predictive mind; dual-process theories; and sequential sampling models of decision-making. I then assessed HECA 's contribution to the control problem's solution and the remaining challenges it confronts. First, I showed that HECA makes some progress on the problem of coherence by positing hierarchical control structures: the higher-level control structures help promote the coherence of its lower-level mechanisms (which can be control structures themselves) through modulating/biasing their response-selection. I conceded, however, that HECA's solution is only partial: it is unclear how the highest-level control structures can make coherent decisions reliably without the modulation of other control structures.

Second, I also demonstrated that HECA makes some progress on the problem of intelligence. The Cooperative Decision Thesis illustrated how epistemic cooperation among heterogeneous information processes can generate more reliable decisions. However, HECA's solution to the problem of intelligence is also incomplete because the challenges of managing dynamics, capacity, and accuracy that are inherent to dynamical decision-making processes in a hierarchically-structured architecture have not been dealt with satisfactorily. Then, I showed that one of the common strategies for addressing these challenges is to assess or predict the accuracy of DVs. Consequently, I drew a few positive lessons from HECA. First, to solve the problem of coherence, some kind of centralized controllers may be necessary. Second, to solve the problem of intelligence, additional control processes need to be implemented in order to overcome the challenges of managing dynamics, capacity, and accuracy.

Part III developed my positive account of neurodemocracy, a hybrid solution of centralized and distributed control with a simple message-passing strategy. I began Chapter 6 by arguing that the BG constitute a central control mechanism that utilizes a simple message-passing strategy. The BG constitute a "central" controller because they exhibit important features of centralized control.

First, the BG systematically integrate a large scale of information from both cortical and subcortical areas through their wide and direct connections. Second, the BG make decisions in a wide range of domains. Finally, the BG also enable robust reinforcement learning in a wide range of domains. In fact, compared to Fodorian central systems, BG integrate a wider range of information from cortical and subcortical areas, make decisions, and facilitate learning for a wider range of domains.

Moreover, unlike the Fodorian central system, the BG are not a rich message-passing controller. They do not rely on complex internal models, the sophisticated transformation of general-purpose representations, or control signals with rich content to coordinate other mechanisms. In fact, BG neither specify the response options nor evaluate them—they rely, instead, on other cognitive mechanisms for action-specification and action-evaluation. The BG select responses based on the accumulation of DVs from action-evaluative information processes, and implement Bayesian inference to select the best option. Afterward, the BG gate the response options through simple control signals of inhibition and disinhibition.

Importantly, the BG controllers are not micro-managers. Through reinforcement learning, the BG assess the reliability of action-evaluating processes in order to improve action-selection in novel contexts. At the same time, the BG train the distributed controllers to select the appropriate actions in these contexts, and disengage when action-selection become habitual.

The neurodemocracy account I have argued for in this thesis makes further progress on the problems of coherence and intelligence. It handles the problem of coherence by positing central controllers that integrate a large scope of information to (learn to) select coherent responses for a wide range of domains. Neurodemocracy also makes progress on the problem of intelligence. The BG controller, through reinforcement learning, improves its assessment of the accuracy of DVs from different information processes. A better assessment, in turn, helps cope with the challenges of managing accuracy, as well as dynamics and capacity.

However, in providing an adequate solution to the problem of intelligence, my account faces some remaining challenges. These owe to the fact that reinforcement learning in the BG controller can only improve its assessment of accuracy by having extensive experience with a given context. As a result, the assessment of accuracy in the novel contexts remains highly problematic. In the next section, I suggest some possible "tricks" the BG and other neural mechanisms might implement to address this remaining challenge.

2. The Epistemic Values of Novelty, Error, and Conflict in Neurodemocracy: Some Speculations

In this section, I want to make some suggestions regarding the "tricks" that the BG implement to assess the accuracy of DVs in novel contexts. These suggestions will, in turn, help address the challenges of managing speed, capacity, and accuracy in novel contexts. As we have seen, these challenges are important for solving the problem of intelligence. In the last chapter, I argued that the BG's function is that of performing a form of Bayesian inference, the multi-hypothesis sequential probability ratio test (SPRT). Subsequent to this, I drew an analogy between the BG's information processing and democratic procedures. In this section, however, my starting point is not Bayesianism, but social choice theory. That is, I will directly examine the epistemic foundation of democracy to borrow inspiration for the BG's computational theory.

I will begin by discussing the Condorcet Jury Theorem, which is often taken to be the epistemic justification for democracy, as a rational model in an idealized condition. Then, I will consider realistic constraints for the cognitive system and modify this rational model to meet them. This

discussion will take us through the familiar issues of speed–accuracy and capacity–accuracy tradeoffs. It will also allow me to introduce models that aim to meet the challenge of assessing the accuracy of accumulated DVs in novel contexts where the estimation of accuracy based on generalization from past experience is likely inadequate. More specifically, these models utilize information about novelty, error, and conflict as signs indicating that the DVs are low in accuracy. In other words, information about novelty, error, and conflict have epistemic value in neurodemocracy insofar as it serves as a fallible sign of unreliable collective decisions. After the discussion of rational models, I will consider how the BG may implement these models and provide some suggestive empirical support that the BG do in fact implement them.

2.1. The Condorcet Jury Theorem as a Rational Model

The Condorcet Jury Theorem, proven formally in the 18th century by Marquis de Condorcet, provides an epistemic justification for majority rule. Majority rule is to be preferred because it is an (imperfect) truth-tracking procedure that, under the right conditions, can produce more reliable decisions than individual members of a group can (List, 2013; List & Goodin, 2001; Weirich, 2013). For example, a trial jury, in which every juror has the common goal to convict the guilty and acquit the innocent, are more likely to give the correct verdict than any one individual juror can, under certain conditions. In other words, the Condorcet Jury Theorem shows that collective intelligence can emerge by aggregating the judgments of less intelligent individuals.

The Condorcet Jury Theorem assumes the following conditions:

(1) **Common Epistemic Goal**: a group (and each member of the group) aim to make a factually correct decision about a state of the world (which is independent of the decision process itself). For example, the trial jury and its members aim to reach a correct verdict about the suspect.

(2) **Two Alternatives**: there are only two possible answers to the question. For example, the suspect is either guilty or not guilty.

(3) **Competence:** each member is better than random at making the correct decision (i.e., the probability of making the correct decision is higher than 0.5). For example, each juror is more competent than a fair coin in making the relevant decision.

(4) **Equal Competence:** each member has the same probability p of making the correct decision.

(5) **Independence:** the decisions of different members are probabilistically independent of each other.[135]

Under these conditions, the Condorcet Jury Theorem proves that the probability of a correct majority decision is greater than the probability of a correct individual decision. Additionally, the probability of a correct majority decision converges to 1, as the number of the members of the group increases. The beneficial epistemic gain of majority rule can be illustrated in Figure 7.1 (below). Also, there are some general trends worth noticing: if we want to increase the reliability of a collective decision, we have two options. First, we can, given a fixed jury size, increase the

[135] I defined the concept of probabilistic independence in footnote 77 in Chapter 4 when I discussed the sequential probability ratio test.

reliability of individual members. Alternatively, if the reliability of individual jurors is better than 0.5, we can increase the number of jurors while fixing their reliability. Of course, we can always do both.

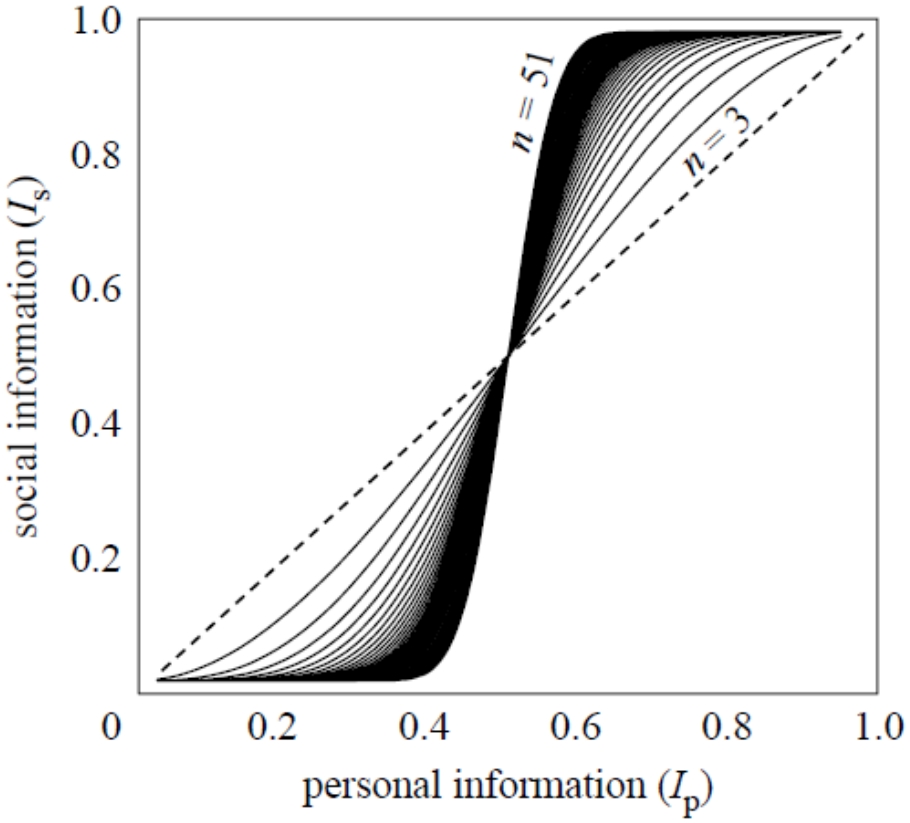

Figure 7.1 relative reliability of collective judgment vs. individual judgment under Condorcet Jury Theorem's assumption. The X-axis represents the reliability of each individual juror, and the Y-axis represents the reliability of the collective judgment. N indicates the size of the jury. For example, if each individual juror has the reliability of 0.6, a jury of 51 will have the reliability of its collective judgment approaching 1. Excerpted from (King & Cowlishaw, 2007).

Meeting these conditions is crucial for the beneficial effect of the majority rule. If the Competence condition is violated, for example, when the jurors are less likely to give the correct answer than to give the incorrect one, the majority decision is in fact less reliable than the individual decision. If the Independence condition is violated— say, because the individual decisions are completely correlated—the majority decision is just as reliable as the individual one. Finally, if there are more than two alternative answers to the question, it is possible that the majority rule will lead to a paradox (List, 2013; Weirich, 2013).

Taken this way, the Condorcet Jury Theorem is an extremely idealized rational model that real-life situations rarely resemble. However, it can be generalized so that the beneficial epistemic gain of majority rule can maintain with the relaxation of some of the above conditions. For example, research shows that the Independence condition, the Competence condition, the Equal Competence condition, and the Two Alternative conditions can be relaxed to some degree without compromising the epistemic gain of majority rule (Hawthorne, 2001; List, 2013; List & Goodin, 2001; Weirich, 2013).

For example, a trial jury in a real-life context tends to have jurors with correlated judgments as they all draw on the same set of evidence. Also, not every juror is competent, nor does their

competences tend to be identical. Nevertheless, the majority rule may still lead to a more reliable decision under these circumstances. Moreover, there are procedures to increase the reliability of the collective decision in more realistic situations, such as weighing the individual decisions according to their reliability.[136] Finally, there are modified collective decision procedures that can avoid the paradox created when there are more than two alternative options (List & Goodin, 2001)—however, this will not concern us because, as we will see in the next section, the model of the BG I consider focuses on making binary yes-or-no decisions.

In short, the Condorcet Jury Theorem demonstrates that when certain conditions obtain, group intelligence emerges non-mysteriously from less intelligent individuals. In terms of its implications for cognitive science, it would not be too inaccurate to think of it as an idealized rational model that illustrates the benefit arising from the pooling of parallel information processing.

Now, the Condorcet Jury Theorem—even with some of its assumptions relaxed—is still too idealized and does not consider real-life constraints, such as time and resources, or their trade-off with accuracy of the decision. With this in mind, let me consider a modified rational model by assuming two conditions. First, many decisions need to be made in parallel, and within limited time. Second, to form a jury for each decision, a fixed group of jurors (limited resources) is drawn that consists of two types of jurors: the higher-level ones that are more reliable but scarce in number and slow, and the lower-level ones that are less reliable but numerous and fast.

Under such circumstances, the rational strategy seems to incorporate the following procedures in the rational model: First of all, we should form the jury for a specific decision with only a few (preferably, lower-level) jurors to begin in order to minimize the demand on resources and maximize the speed. This is because if we draw too many jurors from the fixed group for this decision, there will be fewer jurors available for other decisions, and calling on too many jurors (especially, the higher-level ones) will slow down the speed of decision-making.

Then, if the overall reliability of jurors is too low to reach a correct collective decision with the desired reliability, we can implement three procedures to increase the reliability of the decision-making process (see Figure 7.1 for an approximated illustration, as the figure is based on the assumption of equal competence of individual member). First, we can simply increase the number of jurors by calling on more lower-level jurors (given that their reliability is above 0.5); second, we can increase the overall reliability of jurors by calling on more higher-level jurors; and finally, we can do both. Note that by increasing the reliability of the collective decision, we also increase the demand on resources and time needed to make a decision.

In short, under the constraints of time and resources, we can achieve the right speed-accuracy and resource-accuracy trade-offs by setting a desired reliability for the collective decision-making process, and improve its reliability by recruiting more lower-level and/or higher-level jurors on a need basis.

Finally, we need to consider a rational model under the constraint of unknown reliability. In the last two rational models, in ideal conditions and under time and resource constraints, we have perfect information about the reliability of individual jurors. However, in real life settings, we almost never have such information available as the jurors do not come with their reliability written on their

---

[136] The reliability of collective decision can be optimized by weighing individual decision, where individual *i* with the reliability of *Pi* is given a weight *Wi* proportional to $\log \frac{Pi}{1-Pi}$ ("Condorcet jury theorem," n.d.).

faces. Similarly, considering cognitive science's application of social choice theory, the reliability of a particular information process is usually uncertain, and this uncertainty is especially high in novel situations. We thus need to further adapt the rational model to make it applicable to more realistic situations.

Assume a situation where the exact reliability of each juror is unknown. The rational strategy seems to be to estimate the reliability of either the individual jurors or the entire jury; then, if necessary, apply the procedures discussed in the second rational model under time and resource constraints in order to meet the desired reliability for the collective decision. So, the key question is how to estimate the reliability of individual jurors and the jury as a group. There are many possible (and fallible) procedures; but I will discuss just three of them here, because they are potentially implemented in the BG.

First, the reliability of a juror for a particular decision can be predicted based on their past performance. This is intuitively a rational procedure. For example, if a physicist has provided correct answers to questions in a particular field for years, it is reasonable to believe that the likelihood of her providing a correct answer to the next question in that field is extremely high (and low if her answers have been incorrect). While this procedure can provide a good estimation of the reliability of a juror, it is fairly limited: the prediction of a juror's reliability is only reliable in the same or similar situations in which she has a demonstrated track record. For example, it is hard to assess the likelihood of the physicist providing correct answers in a field, say psychology, in which she has no track record.

Second, reliability can be predicted by looking for signs that have no direct link with past performance, but nevertheless may correlate with the reliability of a juror in the current context of decision-making. These signs include, for example, the presence of novel problems/contexts or evidence of recent errors. It seems commonsensical that signs that a consultant has no training in a problem (i.e., it is a novel problem for him) should lead us to doubt his reliability with regard to this problem. It also does not seem unreasonable that signs that a consultant has made mistakes in the recent past should lead us to lower our prediction of his reliability on a particular problem of our interests.

Finally, the most interesting aspect of this analysis is that we can estimate the reliability of the entire jury directly without estimating the reliability of individual jurors. A means of estimation is to look at the conflict of individual judgments, because they carry information about the reliability of the entire group. If the conflict is very strong (e.g., half of the jurors vote yes, while the other half vote no), the overall reliability is likely poor. For example, in the climate change debates, the presence of conflict in opinions among experts can lead us to doubt the overall reliability of this field, because the opposing sides certainly cannot be both right. In fact, the rationality of this intuition can be proven formally for the jury with non-independent jurors (Bovens & Hartmann, 2004, Chapter 3).

Note that conflicts come to positively correlate with the lower overall reliability of a group (and hence, the lower reliability of the collective decision) only if its members' individual decisions are not completely independent from each other (List, 2004; Weirich, 2013). When they are independent, the conflict of individual judgments does not carry information about the reliability of the jury. For example, the reliability of collective decision in the Condorcet Jury Theorem (with all of its original assumptions met) depends only on the margin of votes (i.e., the difference between the yes and no votes). As a result, a highly coherent majority vote (say, 10 vs. 0) should inspire the same confidence in the reliability of the final decision as a highly incoherent majority vote (say, 55 vs. 45) does (everything else being equal).

In short, novelty, error, and conflict are epistemically valuable because they carry information about the (potentially) lower reliability of individual judgments and, as a result, the lower reliability of the collective judgment as well. Through these indirect measures, we can estimate whether the overall reliability of the jury is lower than desired. Then, we can adopt the same procedures as discussed in the second rational model to improve the collective decision as needed.

To sum up the discussion of the rational models so far, we've learned that democratic procedure (in particular, the majority rule) can increase the reliability of collective decision-making, but it only does so under certain conditions. Moreover, under realistic constraints of limited time and resources, as well as uncertainty about reliability, modified rational models can incorporate procedures that can maintain (approximately and fallibly) the enabling conditions for the emergence of collective intelligence. Most interestingly, they can take advantage of information about novelty, error, and conflict in democratic decision-making in order to make adjustments that help produce a more reliable collective decision. In the following section, I will draw a parallel between the rational models and the basal ganglia mechanisms in order to show, qualitatively, how these rational models can be implemented in the BG.

2.2. Implementing the Rational Models in the Basil Ganglia

Before I discuss the implementation of these rational models in the BG, I need to introduce some relevant pathways and nuclei in the BG and their functions. While I have discussed the BG's structures in Section 2.1 of the last chapter, I did so using Gurney's model. The following discussion is based on Michael Frank's model, another prominent computational model of the BG (O'Reilly & Frank, 2006; Wiecki & Frank, 2013). Frank's model and Gurney's model differ in the functions they assign to the indirect and hyperdirect pathways, as well as in the target phenomena they aim to explain. At this moment, it is unclear which model is empirically better supported, and they may capture complementary aspects of the BG's functioning.

Frank's BG Model can be illustrated in Figure 6.2 (p.120). First, the direct pathway's net effect promotes the selection of a particular action option. Roughly, we can think of it as a pathway aggregating the "yes" votes from the jury (i.e., DVs representing positive evaluation from action-evaluating information processes). Second, the indirect pathway's net effect is discouraging the selection of a particular action option. Roughly, we can think of this as a pathway aggregating the "no" votes from the jury (i.e., DVs representing negative evaluations from action-evaluating information processes). These DVs are aggregated in the BG's output nuclei GPi/SNr. Finally, the hyperdirect pathway, whose net effect is discouraging the selection of *all* action options in the same response-selection process. Roughly, we can think of this pathway as adjusting the threshold for selection. Moreover, action-evaluating information processes at different levels function as "jurors" with different speed, capacity, and reliability. These information processes (jurors) send their DVs (votes) to the BG. Importantly different from Gurney's model, the channels corresponding to the same response option in these three different pathways do not receive identical neural activations in Frank's model.

A few comments are in order before we move on. First, Frank's model focuses on the selection of a particular option (e.g., whether to turn left or not). In doing so, it brackets the issue of how multiple response options are evaluated against each other in response-selection (i.e., deciding between turning left, turning right, and walking straight), which is the focus of Gurney's model. The selection of a particular option is almost always embedded within the evaluation of multiple options in response-selection. This is why the two BG models may be complementary. However, to address how they relate will take us beyond the scope of our current discussion, and I don't think it will affect my main point here. Second, it should be obvious that the Condorcet Jury Theorem

needs to be further generalized to be applicable to cognition: our evaluative mechanisms do not "vote" discretely on "yes" or "no," but provide evaluative signals that represent continuous values. There is research done on how to aggregate optimally evaluative signals of this nature (Lyon & Pacuit, 2013; Yeung, 2014), but I will also have to bracket it for the purposes of my present discussion.

With this in mind, I will begin by considering how the BG implement the idealized rational model of the Condorcet Jury Theorem with some of the Theorem's assumptions relaxed (specifically, the Competence, Equal Competence, and Independence conditions) and the Two Alternatives condition generalized (i.e., instead of yes or no, positive and negative DVs of continuous values). As I discussed earlier, the Condorcet Jury theorem shows us that, under the right conditions, majority rule can lead to a more reliable collective decision. BG can implement a generalized form of this majority rule procedure. The direct and indirect pathways can be seen as collecting "yes" and "no" votes respectively (the positive and negative DVs) to be aggregated at BG's output nuclei, SNr/GPi, where the collective decision is made based on the result of aggregation. In short, the BG seem to take advantage of the "wisdom-of-the-crowd-effect" by implementing the generalized versions of the Condorcet Jury Theorem to increase the reliability of decision-making.

Having addressed the implementation of the generalized Condorcet Jury Theorem, I will now consider the implementation the rational model under the constraints of time and resource. As I discussed earlier, the rational strategy here will be to start with some small number of (preferably, lower-level) jurors; then, if necessary, recruit more lower-level or even higher-level jurors, in order to increase the reliability of the collective decision.

This strategy can be realized in the BG through flexible adjustment of the decision threshold and recruitment of additional evaluative information. The decision threshold, by determining the minimum amount of DVs needed to reach a decision, also determines the desired reliability of the decision. In other words, the higher the threshold is, the higher the final reliability will be (when other conditions are held constant).

Instead of adjusting the threshold directly, most mechanisms implement indirect measures. For example, there is empirical evidence suggesting that the STN implements one of the mechanisms that adjusts the threshold. The STN can send a strong activation through hyperdirect pathway, with the net effect of discouraging the selection of all action options. As a result of this, a stronger direct pathway activation (accumulating DVs reflecting positive evaluations) is needed for a particular response option to be selected. This procedure is functionally equivalent to raising the decision threshold (Ratcliff & Frank, 2012; Shadlen & Kiani, 2013). There is also evidence suggesting that the flexible adjustment of the decision threshold can be implemented by increasing or decreasing the gain of either the DVs or the "urgency" signals (DVs that do not reflect the evaluation of the action option, but simply grow over time to bring the accumulated DVs closer to the decision threshold) (Thura, Cos, Trung, & Cisek, 2014). I will suggest a few more mechanisms shortly when I address the implementation of rational models under the constraints of unknown reliability.

If the desired reliability of a decision is reached quickly (i.e., when DVs reach the decision threshold), a decision can be made with a small set of action-evaluating information processes. However, if the desired reliability is not met, more lower-level or higher-level information processes can be recruited to evaluate action options. The recruitment of more action-evaluating information processes may happen without the help of specific recruiting mechanisms, simply as a consequence of an unmet decision threshold. It is also empirically plausible that there are mechanisms, implemented in the BG or otherwise, that function to recruit information processes in response to the demand for higher reliability; further investigating this possibility is, however,

outside the scope of this thesis. Whichever way these procedures are implemented, they will require more time and resources, but will likely produce a decision that is more reliable.

In short, the flexible adjustment of the decision threshold and recruitment of additional DVs determines the speed–accuracy and resource–accuracy tradeoffs, and constitutes a rational procedure under time and resource constraints. In fact, this rational model may be equivalent to a more complicated version of the sequential probability ratio test (discussed in Chapter 4) that flexibly adjusts its decision threshold.

Finally, I will turn to the implementation of the rational model under the constraints of unknown reliability. This rational model needs to be implemented by the cognitive system in general, because decision-making processes in all domains, even perceptual decision-making, need to constantly assess the reliability of its information processes (Deneve, 2012). However, the need to assess reliability, as well the difficulty of doing so, is especially high in the BG, because of their specialization in novel decision-making. My hunch is that the BG central controllers employ much more complex decision-making and learning mechanisms compared to those of distributed controllers because of this reason.[137] In fact, many features of the BG central controllers can be viewed as implementing smart and fallible tricks that assess and predict the reliability of the information processes, including ones taking advantage of the epistemic values of novelty, error, and conflict in neurodemocracy.

As discussed earlier, the rational strategy under the constraint of unknown reliability is to estimate the reliability of the jury, and improve it on a need basis. In the following, I will discuss the potential implementation of the three different strategies discussed in the last section.

The first strategy, when applied to decision-making in the cognitive system, is to estimate the reliability of information processes based on their past performance. Additionally, with the information about the estimated reliability at hand, weighing the DVs provided by information processes according to their reliability can further increase the reliability of response-selection. This procedure can be implemented by the reinforcement learning (RL) mechanism in the BG as I discussed in the last chapter. Roughly, RL can adjust the sensitivity of the BG's direct and indirect pathways to different sources of inputs based on the reward prediction errors. If an action leads to more rewards than expected, RL will strengthen the associated cortico-striatum connections in the direct pathway, which leads to the associated DVs having a higher gain in the next round of action-selection. We can interpret this increased sensitivity as the higher reliability attributed to the information processes that are providing the DVs. This is because these information processes have made the right decision to promote the selection of a rewarding action. On the other hand, RL will weaken the associated cortico-striatum connections in the indirect pathway, which leads to the associated inputs having lower gain. This is because the information processes providing DVs to the indirect pathway made the wrong decision in discouraging the selection of a rewarding action. Because the direct pathway promotes the selection of an action option, while the indirect pathway discourages the selection of an action option, the net effect leads to a higher probability of the selection of this action in similar situations (presumably a correct decision) next time.

The second strategy, when applied to the cognitive system, is to look for signs indicating that the system is facing a novel context or that it has recently made some errors. Because these signs may

---

[137] That is, in addition to those reasons discussed in Chapter 6, such as the need to perform response-selection reliably with a large number of action options of similar accumulated DVs.

correlate with the lower reliability of the information processes, increasing the decision-threshold and recruiting additional information processes can help achieve the desirable reliability of the final decision. This strategy may also be implemented in the BG. There is empirical research suggesting that the hyperdirect pathway carries information about error and information about surprise (unexpected events), which may indicate novel contexts. This information is computed by other cortical mechanisms (potentially in the dorsal anterior cingulate cortex, right inferior frontal cortex, or pre-supplementary motor area) (Egner, 2011; Shenhav, Cohen, & Botvinick, 2016; Wessel & Aron, 2017). As the activation of the hyperdirect pathway discourages the selection of all competing response options, these signals of novelty and error, carried via the hyperdirect pathway, effectively increase the decision threshold and help achieve the desired reliability of response-selection.

Finally, I think the most interesting aspect of my proposal concerns the third strategy: the use of information about the conflict between information processes as a sign of their lower overall reliability. Again, increasing the threshold and recruiting more information processes can help achieve the desired reliability for decision-making. Several mechanisms in the BG can utilize the conflict of DVs from the direct and indirect pathways to disproportionally increase the DVs from the direct pathway (representing positive evaluation) that are needed to pass the decision threshold. One mechanism involves mutual inhibition between striatal neurons in the direct and indirect pathways (i.e., D1 and D2 neurons). The inhibitory connections from the D2 neurons in the indirect pathways to the D1 neurons in the direct pathways are stronger than the inhibitory connections from the D1 neurons to D2 neurons (Planert, Szydlowski, Hjorth, Grillner, & Silberberg, 2010; Taverna, Ilijic, & Surmeier, 2008). As a result, to cancel out the effect of a small increase of DVs in the indirect pathway, a much larger amount of DVs in the direct pathway are required. The result is equivalent to raising the decision threshold in response to conflicts between DVs. Alternatively, the same effect may also be achieved through feedback from the GPe back to the striatum.[138] The need to extract information about conflicts may help explain why the BG, in contrast to cortical decision-making mechanisms, implement two distinct pathways to accumulate DVs that reflect positive and negative evaluations of the same response option.

In this section, I have suggested several ways in which rational models may be implemented in the BG central controllers. The BG central controllers manage the wisdom-of-the-crowd effect in order to make the best guess at the optimal response in the novel situation. They do so under time and resource constraints as well as a lack of precise information about the accuracy of DVs. One key strategy involves estimating the reliability of information processes, individually or as a whole, through various smart but fallible "tricks," such as those utilizing information of novelty, error, and conflict. This allows the BG to then use information about the accuracy of DVs to help tackle the crucial challenges of managing dynamics, capacity, and accuracy discussed in Chapter 5 and Chapter 6. The BG's capacity to make reliable decisions without being given precise information about the accuracy of the DVs may be one of the reasons why they are so important for decision-making in novel contexts.

In conclusion, I've suggested the fruitfulness of theoretical research that builds rational models based on social choice theory, as well as the empirical work of testing mechanistic hypotheses that implement "neurodemocracy." I have no doubt that there are richer details to these rational models, as well as alternative means for implementing them in the basal ganglia or other neural mechanisms.

---

[138] This point was suggested to me by Bernard Balleine in personal communication.

However, I believe that this speculative discussion has offered a novel neurodemocratic framework with which to think about the basal ganglia and how they address the control problem.

Control is an essential task for human cognitive systems. This thesis has focused on how neural mechanisms address the two important sub-problems of coherence and intelligence, which concern the decision-making side of the control problem. Although what I have offered is far from a complete solution to the control problem, it is my hope that this thesis, as a whole, has made one step toward a better understanding of how the human cognitive system, including its neural, bodily, and potentially environmental substrates, co-ordinates its component mechanisms and utilizes its information flexibly and selectively in order to produce intelligent behavior.

References

Allen, C. (2009). Teleological notions in biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009). Stanford: Metaphysics Research Lab, Stanford University.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, Mass: Harvard University Press.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, N.J: L. Erlbaum Associates.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?*. Oxford; New York: Oxford University Press.

Anderson, J. R., & Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral and Brain Sciences*, *26*(5), 587–601.

Anderson, M. L. (2014). *After phrenology: neural reuse and the interactive brain*. Cambridge, Massachusetts: The MIT Press.

Anderson, M. L. (2015). Précis of after phrenology: neural reuse and the interactive brain. *Behavioral and Brain Sciences*, 1–22.

Arpaly, N., & Schroeder, T. (2014). *In praise of desire*. Oxford; New York: Oxford University Press.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(3), 632–656.

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, *14*(5), 208–215. https://doi.org/10.1016/j.tics.2010.02.001

Atkinson, A. P., & Wheeler, M. (2004). The grain of domains: the evolutionary-psychological case against domain-general cognition. *Mind & Language*, *19*(2), 147–176.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge; New York: Cambridge University Press.

Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford; New York: Oxford University Press.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200.

Badre, D., & Frank, M. J. (2011). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. *Cerebral Cortex*, *22*(3), 527–536.

Bak, I. J., Markham, C. H., & Morgan, E. S. (1981). A striato-striatal connection in rats. In *Soc. Neurosci. Abstr* (Vol. 7, p. 64).

Balleine, B. W., Dezfouli, A., Ito, M., & Doya, K. (2015). Hierarchical control of goal-directed action in the cortical–basal ganglia network. *Current Opinion in Behavioral Sciences*, *5*, 1–7. https://doi.org/10.1016/j.cobeha.2015.06.001

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4–5), 407–419. https://doi.org/10.1016/S0028-3908(98)00033-1

Balleine, B. W., Liljeholm, M., & Ostlund, S. B. (2009). The integrative function of the basal ganglia in instrumental conditioning. *Behavioural Brain Research*, *199*(1), 43–52. https://doi.org/10.1016/j.bbr.2008.10.034

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.). (1995). *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press, USA.

Baron-Cohen, S. (1995). *Mindblindness: As essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Barrett, H. C. (2005a). Enzymatic computation and cognitive modularity. *Mind & Language*, *20*(3), 259–287.

Barrett, H. C. (2005b). Modularity and design reincarnation. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Structure and Contents* (pp. 199–217). Cambridge, MA: Oxford University Press.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological Review*, *113*(3), 628–647. https://doi.org/10.1037/0033-295X.113.3.628

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, *113*(18), 4900–4908. https://doi.org/10.1073/pnas.1520084113

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. MA: Taylor & Francis.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4*(3), 91–99.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(02), 227–287.

Block, N. (2005). Review of Alva Noë, action in perception. *Journal of Philosophy*, *102*(5), 259–272.

Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118–125. https://doi.org/10.1016/j.tics.2006.12.006

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*(2), 442–477.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1655–1670. https://doi.org/10.1098/rstb.2007.2059

Bond, A. H. (2004). An information-processing analysis of the functional architecture of the primate neocortex. *Journal of Theoretical Biology*, *227*(1), 51–79. https://doi.org/10.1016/j.jtbi.2003.10.005

Botvinick, M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208. https://doi.org/10.1016/j.tics.2008.02.009

Botvinick, M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008

Botvinick, M., & Cohen, J. D. (2014). The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers. *Cognitive Science*, *38*(6), 1249–1285. https://doi.org/10.1111/cogs.12126

Botvinick, M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130480–20130480. https://doi.org/10.1098/rstb.2013.0480

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford : Oxford ; New York: Clarendon ; Oxford University Press.

Boyd, R., & Richerson, P. J. (2007). Culture, Adaptation, and Innateness. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Volume 2: Culture and Cognition* (1st ed., pp. 23–38). Cambridge, MA: Oxford University Press.

Brooks, R. A. (1991a). Intelligence without representation. *Artificial Intelligence*, *47*(1), 139–159.

Brooks, R. A. (1991b). New approaches to robotics. *Science*, *253*(5025), 1227–1232.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2), 96–107.

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 133–154). Malden, MA: Blackwell Pub.

Busemeyer, J. R., & Johnson, J. G. (2008). Microprocess models of decision making. *Cambridge Handbook of Computational Psychology*, 302–321.

Buss, D. M. (Ed.). (2005). *The handbook of evolutionary psychology*. Hoboken, N.J: John Wiley & Sons.

Buss, D. M. (Ed.). (2016). *The handbook of evolutionary psychology* (2nd edition). Hoboken, New Jersey: John Wiley & Sons, Inc.

Byrne, M. D. (2012). Unified theories of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(4), 431–438. https://doi.org/10.1002/wcs.1180

Carruthers, P. (2003). On Fodor's problem. *Mind & Language*, *18*(5), 502–523. https://doi.org/10.1111/1468-0017.00240

Carruthers, P. (2006). *The Architecture of the Mind*. Cambridge, MA: Oxford University Press.

Carruthers, P. (2007). Simple heuristics meet massive modularity. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Volume 2: Culture and Cognition* (pp. 181–198). Cambridge, MA: Oxford University Press, USA.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, Mass: MIT Press.

Cherniak, C. (1990). *Minimal Rationality. Cambridge, MA: Bradford*. MIT Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: M.I.T. Press.

Chomsky, N. (1975). *Reflections on language* (1st ed). New York: Pantheon Books.

Chomsky, N. (2005). *Rules and representations*. New York: Columbia University Press.

Chow, S. (2015). Many meanings of "heuristic." *The British Journal for the Philosophy of Science*, *66*(4), 977–1016. https://doi.org/10.1093/bjps/axu028

Chow, S. (2016). Fodor on global cognition and scientific inference. *Philosophical Psychology*, *29*(2), 157–178. https://doi.org/10.1080/09515089.2015.1013208

Churchland, P. M. (1989). *A neurocomputational perspective: the nature of mind and the structure of science*. Cambridge, MA: MIT Press.

Churchland, P. S., & Suhler, C. L. (2014). Agency and Control: The Subcortical Role in Good Decisions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: Free Will and Moral Responsibility* (Vols. 1–4, pp. 309–326). Cambridge, MA: A Bradford Book.

Cisek, P. (2012a). Cortical mechanisms of action selection: the affordance competition hypothesis. In A. K. Seth, T. J. Prescott, & J. J. Bryson (Eds.), *Modelling Natural Action Selectionn*. Cambridge: Cambridge University Press.

Cisek, P. (2012b). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, *22*(6), 927–936. https://doi.org/10.1016/j.conb.2012.05.007

Cisek, P., & Kalaska, J. F. (2010). Neural Mechanisms for Interacting with a World Full of Action Choices. *Annual Review of Neuroscience*, *33*(1), 269–298. https://doi.org/10.1146/annurev.neuro.051508.135409

Cisek, P., & Pastor-Bernier, A. (2014). On the challenges and mechanisms of embodied decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130479–20130479. https://doi.org/10.1098/rstb.2013.0479

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. Cambridge, MA: The MIT Press.

Clark, A. (2001). *Mindware: an introduction to the philosophy of cognitive science*. Oxford: Oxford University Press.

Clark, A. (2007). Soft Selves and Ecological Control. In D. Ross, D. Spurrett, H. Kincaid, & G. L. Stephens (Eds.), *Distributed cognition and the will: individual volition and social context* (pp. 101–122). Cambridge, MA: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204.

Clarke, M. (2004). *Reconstructing reason and representation*. Cambridge, MA: MIT Press.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(3), 332.

Collins, J. (2005). On the input problem for massive modularity. *Minds and Machines*, *15*(1), 1–22. https://doi.org/10.1007/s11023-004-1346-5

Condorcet jury theorem. (n.d.). In *Encyclopedia of Mathematics*. Retrieved from https://www.encyclopediaofmath.org/index.php/Condorcet_jury_theorem

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. BarKow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: new perspectives on cognition and motivation. *Annual Review of Psychology*, *64*(1), 201–229. https://doi.org/10.1146/annurev.psych.121208.131628

Cowie, F. (1998). *What's Within?: Nativism Reconsidered*. New York: Oxford University Press.

Craver, C. F. (2007). *Explaining the brain: mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon Press.

Craver, C., & Tabery, J. (2017). Mechanisms in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017). Stanford: Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/science-mechanisms/

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130478–20130478. https://doi.org/10.1098/rstb.2013.0478

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, *22*(6), 1068–1074. https://doi.org/10.1016/j.conb.2012.05.011

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

Deise, E. C. (2008). *Frame Problems, Fodor's Challenge, and Practical Reason (Doctoral Dissertation)*. University of Maryland, College Park.

Deneve, S. (2012). Making Decisions with Unknown Sensory Reliability. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00075

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown and Co.

Dennett, D. C. (2007). My Body Has a Mind of Its Own. In D. Ross, D. Spurrett, H. Kincaid, & G. L. Stephens (Eds.), *Distributed Cognition and the Will: Individual Volition and Social Context* (1st ed., pp. 93–100). Cambridge, Mass: MIT Press.

Descartes, R. (1998). *Discourse on Method*. (D. A. Cress, Trans.). Hackett Publishing.

Descartes, R. (1984). *The philosophical writings of Descartes*. (J. Cottingham, R. Stoothoff, & D. Murdoch, Trans.) (Vol. 1). New York: Cambridge University Press.

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

DeWolf, T., & Eliasmith, C. (2013). A neural model of the development of expertise. In *Proceedings of the 12th international conference on cognitive modelling* (pp. 119–124). Retrieved from http://iccm-conference.org/2013-proceedings/papers/0020/paper0020.pdf

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dietrich, E., & Fields, C. (1995). The role of the frame problem in Fodor's modularity thesis: a case study of rationalist cognitive science. *Journal of Experimental & Theoretical Artificial Intelligence*, *7*(3), 279–289. https://doi.org/10.1080/09528139508953813

Ding, L., & Gold, J. I. (2013). The Basal Ganglia's Contributions to Perceptual Decision Making. *Neuron*, *79*(4), 640–649. https://doi.org/10.1016/j.neuron.2013.07.042

Dixon, M. L., & Christoff, K. (2014). The lateral prefrontal cortex and complex value-based learning and decision making. *Neuroscience & Biobehavioral Reviews*, *45*, 9–18. https://doi.org/10.1016/j.neubiorev.2014.04.011

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*(6), 1347–1369.

Dugatkin, L. A. (1992). Sexual Selection and Imitation: Females Copy the Mate Choice of Others. *The American Naturalist*, *139*(6), 1384–1389. https://doi.org/10.2307/2462347

Dum, R. P., Bostan, A. C., & Strick, P. L. (2014). Basal Ganglia and Cerebellar Circuits with the Cerebral Cortex. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The cognitive neurosciences* (Firth edition, pp. 419–434). Cambridge, Massachusetts: The MIT Press.

Egner, T. (2011). Surprise! A unifying model of dorsal anterior cingulate function? *Nature Neuroscience*, *14*(10), 1219–1220. https://doi.org/10.1038/nn.2932

Eliasmith, C. (2013b). *How to build a brain: a neural architecture for biological cognition*. Oxford: Oxford University Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*, *338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Evans, J. S. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33–54). New York: Oxford University Press.

Evans, J. S. B. T., & Frankish, K. (Eds.). (2009). *In two minds: dual processes and beyond*. Oxford ; New York: Oxford University Press.

Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, East Sussex, UK: Psychology Press.

Felleman, D. J., & van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate. *Cerebral Cortex*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1

Fodor, J. A. (1975). *The language of thought*. New York: Crowell.

Fodor, J. A. (1983). *The Modularity of Mind* (Vol. 341). Cambridge, MA: MIT press.

Fodor, J. A. (1998). *In critical condition: polemical essays on cognitive science and the philosophy of mind*. Cambridge, Mass: MIT Press.

Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. New York: Oxford University Press, USA.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass*, *5*(10), 914–926.

Frank, M. J., & Badre, D. (2011). Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. *Cerebral Cortex*, *22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Fuller, T., & Samuels, R. (2014). Scientific inference and ordinary cognition: Fodor on holism and cognitive architecture. *Mind & Language*, *29*(2), 201–237.

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, *8*(4), 143–145.

Gershman, S. J., Daw, N. D., Rabinovich, M. I., Friston, K. J., & Varona, P. (2012). Perception, action and utility: The tangled skein. In *Principles of brain dynamics: Global state interactions* (pp. 293–312). Cambridge, MA: MIT Press.

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In Derek J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62–88). Oxford: Blackwell Pub.

Gigerenzer, G. (2006). Bounded and rational. In R. J. Stainton (Ed.), *Contemporary debates in cognitive science* (Vol. 7, pp. 115–133). Hoboken, NJ: Wiley-Blackwell.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Selten, R. (Eds.). (2002). *Bounded rationality: the adaptive toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox.

Glimcher, P. W., & Fehr, E. (Eds.). (2014). *Neuroeconomics: decision making and the brain* (Second edition). Amsterdam: Boston: Elsevier/AP, Academic Press is an imprint of Elsevier.

Godfrey-Smith, P. (2001). Three kinds of adaptationism. In S. Orzack & E. Sober (Eds.), *Adaptationism and optimality* (pp. 335–357). Cambridge; New York: Cambridge University Press.

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Goldman, M., Compte, A., Wang, X., & Squire, L. R. (2009). *Encyclopedia of Neuroscience*.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*(6), 401–410.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics*, *84*(6), 411–423.

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, *9*(12), 934–946. https://doi.org/10.1038/nrn2497

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814.

Hardcastle, V. G., & Hardcastle, K. (2015). Marr's Levels Revisited: Understanding How Brains Break. *Topics in Cognitive Science*, n/a–n/a. https://doi.org/10.1111/tops.12130

Haugeland, J. (1985). *Artificial Intelligence—the very idea, A Bradford Book*. The MIT Press, Cambridge, Mass.

Hawthorne, J. (2001). Voting in Search of the Public Good: The Probabilistic Logic of Majority Judgements. *University of Oklahoma*. Retrieved from http://faculty-staff.ou.edu/H/James.A.Hawthorne-1/Hawthorne--Jury-Theorems.pdf

Hebb, D. O. (2002). *The Organization of Behavior: A Neuropsychological Theory* (New Ed edition). Mahwah, N.J: Psychology Press.

Hélie, S., Ell, S. W., & Ashby, F. G. (2015). Learning robust cortico-cortical associations with the basal ganglia: An integrative review. *Cortex*, *64*, 123–135. https://doi.org/10.1016/j.cortex.2014.10.011

Heyes, C. M., & Galef, B. G. (Eds.). (1996). *Social learning in animals: the roots of culture*. San Diego: Academic Press.

Hohwy, J. (2013). *The predictive mind* (First edition). Oxford; New York, NY, United States of America: Oxford University Press.

Horgan, T. E., & Tienson, J. (1996). *Connectionism and the Philosophy of Psychology*. Cambridge, MA: The MIT Press.

Huebner, B. (2014). *Macrocognition: a theory of distributed minds and collective intentionality*. Oxford: Oxford University Press.

Humphries, M. D. (2015). Basal Ganglia: Mechanisms for Action Selection. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of Computational Neuroscience* (pp. 351–356). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-6675-8

Hung, T.-W. (2014). Why the Enzyme Model of Modularity Fails to Explain Higher Cognitive Processes. *NTU Philosophical Review*, (48), 159–190.

Hurley, S. (2001). Perception and action: Alternative views. *Synthese*, *129*(1), 3–40.

Ito, M., & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, *21*(3), 368–373. https://doi.org/10.1016/j.conb.2011.04.001

Jackendoff, R. (2002). Review of the book The Mind Doesn't Work That Way: the Scope and Limits of Computational Psychology, by J. A. Fodor. *Language*, *78*(1), 164–170.

Jaeger, D., & Jung, R. (Eds.). (2015). *Encyclopedia of Computational Neuroscience*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-6675-8

Jeffares, B., & Sterelny, K. (2012). Evolutionary Psychology. In E. Margolis, R. Samuels, & S. P. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford University Press.

Kable, J. W., & Glimcher, P. W. (2009). The Neurobiology of Decision: Consensus and Controversy. *Neuron*, *63*(6), 733–745. https://doi.org/10.1016/j.neuron.2009.09.003

Kahneman, D. (2013). *Thinking, fast and slow* (1st pbk. ed). New York: Farrar, Straus and Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: heuristics and biases*. Cambridge; New York: Cambridge University Press.

Kaplan, M. (1981). A Bayesian theory of rational acceptance. *The Journal of Philosophy*, *78*(6), 305–330.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A Hierarchy of Time-Scales and the Brain. *PLoS Computational Biology*, *4*(11), e1000209. https://doi.org/10.1371/journal.pcbi.1000209

King, A. J., & Cowlishaw, G. (2007). When to use social information: the advantage of large group size in individual decision making. *Biology Letters*, *3*(2), 137–139. https://doi.org/10.1098/rsbl.2007.0017

Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., … Montague, P. R. (2015). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*, 201513619. https://doi.org/10.1073/pnas.1513619112

Knowlton, B. (2015). Basal Ganglia: Habit Formation. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of Computational Neuroscience* (pp. 336–349). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-6675-8

Koechlin, E. (2016). Prefrontal executive function and adaptive behavior in complex environments. *Current Opinion in Neurobiology*, *37*, 1–6. https://doi.org/10.1016/j.conb.2015.11.004

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, *302*(5648), 1181–1185. https://doi.org/10.1126/science.1088545

Laland, K. N. (2002). Imitation, Social Learning, and Preparedness as Mechanisms of Bounded Rationality. In Gigerenzer & R. Selten (Eds.), *Bounded rationality: the adaptive toolbox* (pp. 233–248). Cambridge, MA: MIT Press.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural Basis of Reinforcement Learning and Decision Making. *Annual Review of Neuroscience*, *35*(1), 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512

Lieberman, M. D. (2007). The X-and C-Systems: The Neural Basis of Automatic and Controlled Social Cognition. In E. Harmon-Jones & P. Winkielman (Eds.), *Social neuroscience: integrating biological and psychological explanations of social behavior* (pp. 290–315). New York: Guilford Press.

Liljeholm, M., & O'Doherty, J. P. (2012). Contributions of the striatum to learning, motivation, and performance: an associative account. *Trends in Cognitive Sciences*, *16*(9), 467–475. https://doi.org/10.1016/j.tics.2012.07.007

List, C. (2004). On the Significance of the Absolute Margin. *The British Journal for the Philosophy of Science*, *55*(3), 521–544. https://doi.org/10.1093/bjps/55.3.521

List, C. (2013). Social Choice Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013). Stanford: Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2013/entries/social-choice/

List, C., & Goodin, R. E. (2001). Epistemic democracy: generalizing the Condorcet jury theorem. *Journal of Political Philosophy*, *9*(3), 277–306.

Lyon, A., & Pacuit, E. (2013). The Wisdom of crowds: methods of human judgement aggregation. In P. Michelucci (Ed.), *Handbook of Human Computation* (pp. 599–614). New York: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-8806-4_47

Machery, E. (2008). Massive modularity and the flexibility of human cognition. *Mind & Language*, *23*(3), 263–272. https://doi.org/10.1111/j.1468-0017.2008.00341.x

Margolis, E., & Laurence, S. (2012). In defense of nativism. *Philosophical Studies*. https://doi.org/10.1007/s11098-012-9972-x

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. Cambridge, Mass.: MIT Press.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning. *Journal of Neuroscience*, *31*(7), 2700–2705. https://doi.org/10.1523/JNEUROSCI.5499-10.2011

McMillen, T., & Holmes, P. (2006). The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology*, *50*(1), 30–57.

Menary, R. (Ed.). (2010). *The extended mind*. Cambridge, Mass: MIT Press.

Menary, R. (2014). Neural plasticity, neuronal recycling and niche construction. *Mind & Language*, *29*(3), 286–303.

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, *30*(01). https://doi.org/10.1017/S0140525X07000891

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.

Minsky, M. L. (1986). *The society of mind*. New York: Simon and Schuster.

Monsell, S. (1996). Control of mental processes. In V. Bruce (Ed.), *Unsolved mysteries of the mind: Tutorial essays in cognition* (pp. 93–148). East Sussex: Erlbaum.

Moors, A., & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/37/0033-2909.132.2.297

Murphy, D. (2006). On Fodor's analogy: why psychology is like philosophy of science after all. *Mind & Language*, *21*(5), 553–564.

Nelson, A. B., & Kreitzer, A. C. (2014). Reassessing Models of Basal Ganglia Function and Dysfunction. *Annual Review of Neuroscience*, *37*(1), 117–135. https://doi.org/10.1146/annurev-neuro-071013-013916

Newell, A. (1980). Reasoning, problem-solving and decision processes. In R. Nickerson (Ed.), *Attention and Performance VIII* (pp. 693–718). Hillsdale, NJ: Erlbaum.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, Mass: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs NJ: Prentice-Hall.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, *13*(11), 4700–4719.

Oppenheimer, D. M., & Kelso, E. (2015). Information Processing as a Paradigm for Decision Making. *Annual Review of Psychology*, *66*(1), 277–294. https://doi.org/10.1146/annurev-psych-010814-015148

O'Reilly, R. C., & Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*(2), 283–328. https://doi.org/10.1162/089976606775093909

O'Reilly, R. C., Herd, S. A., & Pauli, W. M. (2010). Computational models of cognitive control. *Current Opinion in Neurobiology*, *20*(2), 257–261. https://doi.org/10.1016/j.conb.2010.01.008

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, 0956797612463080.

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The Mixed Instrumental Controller: Using Value of Information to Combine Habitual Choice and Mental Simulation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00092

Picard, N., & Strick, P. L. (1996). Motor areas of the medial wall: a review of their location and functional activation. *Cerebral Cortex*, *6*(3), 342–353.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4

Pinker, S. (1994). *The language instinct* (1st ed). New York: W. Morrow and Co.

Pinker, S. (1999). *How the mind works*. New York: W. W. Norton & Company.

Pinker, S. (2005). So how does the mind work? *Mind & Language*, *20*(1), 1–24.

Planert, H., Szydlowski, S. N., Hjorth, J. J. J., Grillner, S., & Silberberg, G. (2010). Dynamics of Synaptic Transmission between Fast-Spiking Interneurons and Striatal Projection Neurons of the Direct and Indirect Pathways. *Journal of Neuroscience*, *30*(9), 3499–3507. https://doi.org/10.1523/JNEUROSCI.5139-09.2010

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238. https://doi.org/10.1038/22268

Port, R. F., & Van Gelder, T. (Eds.). (1995). *Mind as motion: explorations in the dynamics of cognition*. Cambridge, Mass: MIT Press.

Potts, R. (1996). *Humanity's descent: the consequences of ecological instability* (1st ed). New York: Morrow.

Prescott, T. J. (2007). Forced Moves or Good Tricks in Design Space? Landmarks in the Evolution of Neural Mechanisms for Action Selection. *Adaptive Behavior*, *15*(1), 9–31. https://doi.org/10.1177/1059712306076252

Prescott, T. J., Redgrave, P., & Gurney, K. (1999). Layered Control Architectures in Robots and Vertebrates. *Adaptive Behavior*, *7*(1), 99–127. https://doi.org/10.1177/105971239900700105

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural Computation*, *24*(5), 1186–1229.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Redgrave, P., Gurney, K., Stafford, T., Thirkettle, M., & Lewis, J. (2013). The role of the basal ganglia in discovering novel actions. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems* (pp. 129–150). Berlin; New York: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-32375-1_6

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, *89*(4), 1009–1023. https://doi.org/10.1016/S0306-4522(98)00319-4

Redgrave, P., Vautrelle, N., & Reynolds, J. N. J. (2011). Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neuroscience*, *198*, 138–151. https://doi.org/10.1016/j.neuroscience.2011.07.060

Richardson, R. C. (2010). *Evolutionary Psychology as Maladapted Psychology*. Cambridge, MA: A Bradford Book.

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: how culture transformed human evolution*. Chicago: University of Chicago Press.

Richerson, P. J., & Boyd, R. (2012). Rethinking Paleoanthropology: A World Queerer Than We Supposed. In G. Hatfield & H. Pittman (Eds.), *Evolution of Mind, Brain, and Culture*. Philadelphia, PA: University of Pennsylvania Press.

Robbins, P., & Aydede, M. (Eds.). (2008). *The Cambridge Handbook of Situated Cognition* (1st ed.). Cambridge: Cambridge University Press.

Robert C. Richardson. (1998). Heuristics and satisficing. In W. Bechtel, G. Graham, & D. A. Balota (Eds.), *A companion to cognitive science* (pp. 566–575). Malden, Mass: Blackwell.

Rupert, R. D. (2011). Embodiment, consciousness, and the massively representational mind. *Philosophical Topics*, *39*(1), 99–120.

Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. *The British Journal for the Philosophy of Science*, *49*(4), 575–602.

Samuels, R. (2002). Nativism in cognitive science. *Mind & Language*, *17*(3), 233–265.

Samuels, R. (2005). The Complexity of Cognition. In P. Carruthers, S. Laurence, & S. P. Stich, *The innate mind: structure and contents* (Vol. 1, p. 107). Oxford: Oxford University Press.

Samuels, R. (2006). Is the human mind massively modular. In R. J. Stainton (Ed.), *Contemporary debates in cognitive science, ed. RJ Stainton* (pp. 37–56). Hoboken, NJ: Wiley-Blackwell.

Samuels, R. (2009a). Nativism. In P. Calvo & J. Symons (Eds.), *The Routledge companion to philosophy of psychology*. Abingdon: Taylor & Francis.

Samuels, R. (2009b). The magical number two, plus or minus: Dual process theory as a theory of cognitive kinds. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: dual processes and beyond* (pp. 129–146). Oxford; New York: Oxford University Press.

Samuels, R. (2010). Classical computationalism and the many problems of cognitive relevance. *Studies In History and Philosophy of Science Part A*, *41*(3), 280–293. https://doi.org/16/j.shpsa.2010.07.006

Samuels, R. (2012). Massive modularity. In E. Margolis & S. P. Stich (Eds.), *The Oxford handbook of philosophy of cognitive science* (pp. 60–92). New York: Oxford University Press.

Samuels, R. I. (1998). *Massively modular minds: The nature, plausibility and philosophical implications of evolutionary psychology (doctoral dissertation)*. Rutgers University, New Brunswick. Retrieved from https://philpapers.org/rec/SAMMMM-2

Samuels, R., Stich, S., & Tremoulet, P. D. (1999). Rethinking Rationality. In E. Lepore & Z. Pylyshyn (Eds.), *Rutgers University Invitation to Cognitive Science*. Oxford: Basil Blackwell.

Scholl, B. J. (2005). Innateness and (Bayesian) visual perception. In P. Carruthers, S. Laurence, & S. P. Stich (Eds.), *The innate mind: Structure and contents* (p. 34). New York: Oxford University Press.

Schroeder, T. (2004). *Three faces of desire*. Oxford; New York: Oxford University Press.

Selfridge, O. G. (1989). Pandemonium: a paradigm for learning. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing: foundations of research* (pp. 115–122). Cambridge, MA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=104389

Seth, A. K., Prescott, T. J., & Bryson, J. J. (2012). *Modelling natural action selection*. Cambridge; New York: Cambridge University Press.

Shadlen, M. N., & Kiani, R. (2013). Decision Making as a Window on Cognition. *Neuron*, *80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shanahan, M. (2012). The brain's connective core and its role in animal cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2704–2714. https://doi.org/10.1098/rstb.2012.0128

Shanahan, M. (2016). The Frame Problem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016). Stanford: Metaphysics Research Lab, Stanford University. Retrieved from http://plato.stanford.edu/archives/spr2016/entries/frame-problem/

Shanahan, M., & Baars, B. J. (2005). Applying global workspace theory to the frame problem. *Cognition*, *98*(2), 157–176.

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, *19*(10), 1286–1291. https://doi.org/10.1038/nn.4384

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review; Psychological Review*, *63*(2), 129.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1–20.

Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, *6*(1), 1–10.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). New York: Cambridge University Press.

Sperber, D. (2007). Modularity and relevance. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Volume 2: Culture and Cognition* (1st ed.). New York: Oxford University Press.

Sperber, D., & Hirschfeld, L. (2006). Culture and modularity. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Volume 2: Culture and Cognition* (pp. 149–164). New York: Oxford University Press.

Sterelny, K. (2006). Cognitive load and human decision, or, three ways of rolling the rock up hill. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Volume 2: Culture and Cognition* (pp. 218–233). New York: Oxford University Press.

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA: MIT Press.

Sterelny, K. (2014). Cooperation, culture, and conflict. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axu024

Sterelny, K., & Griffiths, P. E. (1999). *Sex and death: An introduction to philosophy of biology*. Chicago: University of Chicago Press.

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2012). Learning to Select Actions with Spiking Neurons in the Basal Ganglia. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00002

Stocco, A., Lebiere, C., & Anderson, J. R. (2010). Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination. *Psychological Review*, *117*(2), 541–574. https://doi.org/10.1037/a0019077

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*. Retrieved from http://www.nature.com/nrn/journal/vaop/ncurrent/full/nrn3838.html

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, Mass: A Bradford Book.

Taverna, S., Ilijic, E., & Surmeier, D. J. (2008). Recurrent Collateral Connections of Striatal Medium Spiny Neurons Are Disrupted in Models of Parkinson's Disease. *Journal of Neuroscience*, *28*(21), 5504–5512. https://doi.org/10.1523/JNEUROSCI.5493-07.2008

Teodorescu, A. R., & Usher, M. (2013). Disentangling Decision Models: From Independence to Competition. *Psychological Review January 2013*, *120*(1), 1–38. https://doi.org/10.1037/a0030776

Thagard, P. (2001). How to Make Decisions: Coherence, Emotion, and Practical Inference. In E. Millgram (Ed.), *Varieties of practical reasoning* (pp. 355–372). Cambridge, Mass: MIT Press.

Thagard, P. (2012). Cognitive architectures. In K. Frankish & W. Ramsey (Eds.), *The Cambridge handbook of cognitive science* (pp. 50–70). Cambridge: Cambridge University Press.

Thorndike, E. L. (1905). *Elements of psychology*. New York: A. G. Seiler.

Todd, P. M., & Gigerenzer, G. (2012). *Ecological Rationality: Intelligence in the World*. New York: Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195315448.001.0001/acprof-9780195315448

Todd, P. M., Hills, T. T., & Robbins, T. W. (Eds.). (2012). *Cognitive Search: Evolution, Algorithms, and the Brain*. Cambridge, MA: MIT Press.

Tooby, J., & Cosmides, L. (1995a). Foreword. In S. Baron-Cohen & S. Baron-Cohen, *Mindblindness: As essay on autism and theory of mind* (pp. xi–xviii). Cambridge, MA: MIT Press.

Tooby, J., & Cosmides, L. (1995b). The psychological foundations of culture. In J. H. Barkow (Ed.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press, USA.

Tooby, J., & Cosmides, L. (2000). Toward mapping the evolved functional organization of mind and brain. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 1167–1178). Cambridge, Mass: MIT Press.

Tooby, J., & Cosmides, L. (2016). The theoretical foundations of evolutionary psychology. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (Second, pp. 3–87). Hoboken, New Jersey: Wiley. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781119125563.evpsych101/full

Tsetsos, K., Gao, J., McClelland, J. L., & Usher, M. (2012). Using Time-Varying Evidence to Test Models of Decision Dynamics: Bounded Diffusion vs. the Leaky Competing Accumulator Model. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00079

Turner, R. S., & Desmurget, M. (2010). Basal ganglia contributions to motor control: a vigorous tutor. *Current Opinion in Neurobiology*, *20*(6), 704–716. https://doi.org/10.1016/j.conb.2010.08.022

Turner, R. S., & Pasquereau, B. (2014). Basal Ganglia Function. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The cognitive neurosciences* (Firth edition, pp. 435–450). Cambridge, Massachusetts: The MIT Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124.

Uithol, S., van Rooij, I., Bekkering, H., & Haselager, P. (2012). Hierarchies in action and motor control. *Journal of Cognitive Neuroscience*, *24*(5), 1077–1086.

Usher, M., & McClelland, J. L. (2001). The Time Course of Perceptual Choice: The Leaky, Competing Accumulator Model. *Psychological Review July 2001*, *108*(3), 550–592.

Van Essen, D. C., Felleman, D. J., DeYoe, E. A., Olavarria, J., & Knierim, J. (1990). Modular and Hierarchical Organization of Extrastriate Visual Cortex in the Macaque Monkey. *Cold Spring Harbor Symposia on Quantitative Biology*, *55*(0), 679–696. https://doi.org/10.1101/SQB.1990.055.01.064

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.

Varela, F. J., Thompson, E., & Rosch, E. (2016). *The embodied mind: cognitive science and human experience* (Revised Edition). Cambridge, MA: MIT Press.

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, *150*(3811), 563–565.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*(5), 955–968.

Wang, X.-J. (2012). Neural dynamics and circuit mechanisms of decision-making. *Current Opinion in Neurobiology*, *22*(6), 1039–1046. https://doi.org/10.1016/j.conb.2012.08.006

Weirich, P. (2013). Condorcet's Jury Theorem. In *International Encyclopedia of Ethics*. Oxford: Blackwell Publishing Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781444367072.wbiee038/abstract

Weiskopf, D. A. (2002). A critical review of Jerry A. Fodor's The Mind Doesn't Work That Way. *Philosophical Psychology*, *15*(4), 551–562. https://doi.org/10.1080/0951508021000042067

Wessel, J. R., & Aron, A. R. (2017). On the Globality of Motor Suppression: Unexpected Events and Their Influence on Behavior and Cognition. *Neuron*, *93*(2), 259–280. https://doi.org/10.1016/j.neuron.2016.12.013

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, *120*(2), 329–355. https://doi.org/10.1037/a0031542

Wilson, R. A. (2004). What computers (still, still) can't do: Jerry Fodor on computation and modularity. In R. J. Stainton, M. Ezcurdia, & C. Viger (Eds.), *New Essays in Philosophy of Language and Mind*. Calgary: University of Calgary Press.

Wilson, R. A., & Foglia, L. (2011). Embodied Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2011). Stanford: Metaphysics Research Lab, Stanford University. Retrieved from http://plato.stanford.edu/archives/fall2011/entries/embodied-cognition/

Wimsatt, W. C. (1994). The ontology of complex systems: levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy*, *20*(2), 207–274.

Yeung, S. (2014). A hierarchical Bayesian model for improving wisdom of the crowd aggregation of quantities with large between-informant variability. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Retrieved from https://mindmodeling.org/cogsci2014/papers/541/paper541.pdf

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6), 464–476. https://doi.org/10.1038/nrn1919