Automaticity and Control in Human Action

by

Juan Pablo Bermúdez-Rey

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy in Philosophy
Department of Philosophy
University of Toronto

© Copyright by Juan Pablo Bermúdez-Rey 2016
Since a large portion of our behaviour is automatic, what kind of agency and control do we humans have, and what are its limits? I address the question both from a historical perspective (through Aristotle’s philosophy) and a contemporary approach (engaging with recent empirical research).

I begin with Aristotle’s action theory, which has remarkable resonances with current issues. He claims our goals are determined by non-rational habituation, so the question arises how we can control our life’s ends if they cannot be set rationally. I argue that for Aristotle we control our goals insofar as we carefully shape our habits. Care for habits belongs, not just to the individual (who cannot care for her habituation’s first stages), but primarily to the political community. So individual autonomy depends on social care practices; and reason can establish goals not directly, but by carefully shaping habits.

I go on to argue that empirical evidence justifies a broadly similar account.

Intellectualists hold that all actions must be produced through reflection (i.e. a top-down coordination by reference to a goal); anti-intellectualists argue that some actions can be produced without reflection. Anti-intellectualists receive preliminary empirical support from evidence that there is a kind of normativity proper to automatic (i.e. reflection-independent) processes, which reveals the existence of automatic control. Is automatic control sufficient for the production of intentional actions? Anti-intellectualists have often pointed to skilful bodily action to argue that it is, but I contend that even skilled action requires top-down, reflective control, confirming recent intellectualist proposals.
So anti-intellectualism cannot find support in skilled-action phenomena; but it may find more promising evidence in decision-making studies. Framing effects in decision-making reveal that we control intention-formation processes automatically through *background control*: intuitive sense-making processes that are nonetheless susceptible to cognitive biases largely incorrigible by reflection. However, smart environmental design is able to protect intuitive control over decision-making. This supports a thesis called *indirect intellectualism*: like Aristotle, we should conceive of individual control over action as significantly dependent on reflectively shaping our practical environment.
This work is dedicated to the memory of Daniel Ramírez.

La pérdida de su vida abrió un vacío que nunca se llenará. Este trabajo no podría reemplazar las ideas que perdimos cuando lo perdimos a él. Sin embargo, en las horas más oscuras su recuerdo me ha dado fuerzas para no rendirme y seguir andando el camino que por fortuna llegamos a compartir.
Acknowledgements

As the path that leads to this work’s completion comes to an end, I am filled with a sense of overwhelming gratitude. I thank the University of Toronto’s Philosophy Department and its Collaborative Program in Ancient and Medieval Philosophy, for a transformative learning experience whose benefits will continue to bear fruit for the rest of my life. I also thank the admirable human settlement that is the city of Toronto for the many life lessons it has afforded me.

This project benefited enormously from the generosity of people who read, discussed, and commented on it, without whom the final version would lack so much richness and depth. Mark Kingwell, this work’s supervisor, was remarkable in his steadfast commitment to the project and his confidence that, despite all the dead-ends and sudden transformations along the way, everything would make sense in the end. I am enormously thankful for his relentless philosophical engagement with my ideas. I thank Brad Inwood, Evan Thompson, and Wil Cunningham, committee members, for their openness to currents of thought alien to their respective fields of expertise (a *sine qua non* condition for this project’s completion), and for their keen critical eyes and their constant academic support throughout the project’s reincarnations. Brad deserves an especial acknowledgement for having convinced me that the broader scope I wanted to find for my work would be enriched by my background in ancient Greek philosophy. In this—as in most things—he was right! I will attempt to continue this bridge-building endeavour in the future.

Alejandro Rosas was a sort of surrogate supervisor during my time in Bogotá, reading and commenting on drafts of most of the work. I have learned immensely from our conversations and his straightforward approach to philosophy. Our discussions of philosophical psychology provided invaluable help in understanding the literature’s intricacies. I hope that our collaborations extend well into the future. Adelaida Barrera read most of the text (even its most arid parts, for which she is commendable), helping me clarify and enrich crucial aspects of the argument. For this, and so much more, I will remain grateful. Receiving Jennifer Nagel’s comments on an early version of Chapter 3 had a Socratic significance for me: thanks to them, I realized that I did *not* know many things I thought I knew. This was very difficult to come to terms with at the time, but the experience proved invaluable for correcting serious mistakes and reaching a better understanding of things that would have otherwise escaped me. I thank her enormously for that, and for her constant support as the Department’s Graduate Coordinator. Luisa Monsalve and I shared many moments of Ph.D. exhaustion, but both made it through the dissertation, which makes me happy. I thank her for her kindness and trust during my work at the Universidad Externado.
Santiago Arango-Muñoz, Klaus Corcilius, Laura Gómez-Espíndola, and Javier Guillot offered rich and detailed comments on parts of the dissertation. Ricardo Salles and other colleagues from UNAM’s Instituto de Investigaciones Filosóficas and the Universidad del Valle’s Philosophy Department gave valuable comments on the first Chapter. Many of my fellow graduate students at U of T’s Philosophy Department gave extensive feedback in two Grad Forum sessions that were the first times many of these ideas (fragile and immature as they then were) saw the light of academic discussion. Gerard Baron, Hernando Bermúdez, Alessandro Bonello, Paul Jarvey, Natalie Mathieson, Natalie Sánchez, and Cameron Woloshyn, a group of my (mostly non-philosophy) philosophical friends in Toronto, gave me great insights into what became the first Chapter. I feel a large part of my growth in recent years was due to the experiences I had the privilege of sharing with them and my other Toronto friends, Hasko von Kriegstein, Dave Suarez, Mike Blezy, and Marine Gobert. I hope these friendships last indefinitely. Arturo Durán, Juan Pablo Galvis, Andrés Gualdrón, Javier Guillot, Jorge Patiño, Jairo Pérez, Bibiana Rojas, Alejandro Villate are friends who make the meaning of the word ‘home’ palpable every time we get together. I am thankful to all of the people mentioned above for making my life richer and more meaningful.

In this section authors usually thank their partners for having put up with their bad mood, absent-mindedness, and lack of time while in the writing process. This is not my case, because both I and Adelaida were both going through all that while writing our theses simultaneously. The extreme levels of tension and exhaustion that took over our house during the crucial times can be imagined! I want to thank her for having been a constant source of love and hope, an illuminating companion, and also an example of how good writing is done. She knows I admire her writing, as well as many other aspects of her incredible self. I should also thank Eli and Manchis for having come to brighten up our home at precisely the right time.

Finally, I thank my in-laws Fernando Barrera and Berta Daza for their caring support throughout the good and not so good parts of these years; my siblings, Hernando and María Carolina Bermúdez-Rey, for being sources of inspiration and growth throughout my life; and my parents Hernando Bermúdez-Gómez and Constanza Rey-Pinzón, for enabling me to carefully shape my own character, by giving me the right tools at the right times, and always supporting the paths I’ve chosen for my life. Now that this is done, I hope we can all hang out more often! A final thanks to all the anonymous people who have worked to make life in Colombia better throughout these decades, to the point that many have sacrificed their own lives and dreams. I hope I can now add my efforts to theirs in the construction of an increasingly just and inclusive society.
### Table of contents

**Introduction** xii

**Chapter 1: Habit and reason in Aristotle’s theory of agency** 1

- §1.1 Principles of animal agency 3
  - 1.1.1. Aristotle’s vocabulary of action 3
  - 1.1.2. The principles of animal agency 4
- §1.2. Principles of human agency: intellectualism and its limits 9
  - 1.2.1. Intellectualist interpretations of human agency 9
  - 1.2.2. The limitation of intellectualism: goal-setting 13
  - (A) There is no reasoning of starting points 13
  - (B) Deliberation is not about the goals 14
  - (C) Habituation determines the goals 15
  - Replying to some intellectualist objections 18
  - (D) Specification as constitution 20
  - (E) Wish, the rational desire 23
  - (F) Practical nous 29
- §1.3. Habit and the structure of human agency 36
  - 1.3.1. How habituation works 36
  - What is habituation? 38
    - (A) Habit as an extension of nature 38
    - (B) Repetition, endurance, and pleasure 40
  - Habituation and development 43
    - (C) Habituation and moral development 43
    - (D) Habituation and intellectual development 45
  - 1.3.2. The circular structure of human agency 49
- §1.4. Reason as care: A defence of indirect Intellectualism 52
  - 1.4.1. Care for habits as necessary for control 53
  - 1.4.2. Legislation, reason, and care 56
    - Public vs. private care 56
    - Law and reason 58
    - Public vs. private reason 61
    - Why care should be public 63
  - 1.4.3. Indirect Intellectualism 64
    - Indirect Intellectualism, the role of ethics, and the human-animal distinction 66
      - (A) The two theses of indirect Intellectualism 66
      - (B) What the study of ethics is for 67
      - (C) The difference between animals and humans 68
    - Extended control over action 70
- §1.5. Looking ahead 71
Chapter 4: Automaticity and control in skilled action

§4.1. The problem of control in skilled action

4.1.1. A recent history of the debate over skill

How reflection controls automatic associations 141

Three psychological phenomena against intellectualism 143

(A) Reflection, choking, and the yips 143
(B) Expertise-induced amnesia 144
(C) Expert confabulation 145

4.1.2. On the notion of 'skill'

Skill and automaticity 145
Skill and control 146
Kinds of skills 147
Skill and expertise 147

§4.2. Does reflection lead to choking?

4.2.1. Clarifying the terms

Performance failure under pressure: one or many? 148
Choking, The Yips, and reflective control's proper role 150

4.2.2. Do these phenomena support anti-intellectualism about skill? 152

§4.3. Expertise-induced amnesia

4.3.1. The anti-intellectualist case 154

4.3.2. Intellectualist replies

The experiments are largely inconclusive 155
Alternative expert reports suggest otherwise 156

§4.4. Expert confabulation

4.4.1. Evidence for expert confabulation 158

4.4.2. “Expert”? “Confabulation”?

§4.5. Reflection’s role in skilled action

4.5.1. Reflection as the intention-based modulation of attention 161

4.5.2. Fringe actions

Are tying shoelaces and catching flying balls intentional actions? 165
Action, automatic control, and the perils of distraction 166
The possibility of fringe actions 169

Chapter 5: Automaticity and control in decision-making

§5.1. Intentions, decisions, and the problem of autonomy

5.1.1 Basic agency and the problem of autonomy 171
5.1.2. Intention formation and decision-making 172

§5.2. Framing effects on decision-making: The limits of agentive control

5.2.1. Perceptual illusions and cognitive illusions 173
5.2.2. Two cases of cognitive illusions 175
5.2.3. The cognitive mechanisms behind framing effects 176

Framing and valuing 176
A dual-process model of framing effects 178

5.2.4. Framing effects and the limits of autonomy 179
   Framing effects and decision myopia 180
   (A) We let opportunities pass us by 180
   (B) We are bad losers 181
   (C) We are bad cooperators 182
   (D) We lack self-control 183
   Decision myopia can lead to manipulation 184
   (E) A wealth of opportunities for pulling people’s strings 185
   (F) Manipulation and autonomy: Philosophical problems 186
   Decision myopia, manipulation, and autonomy 188

§5.3. Debiasing framing effects 189
   5.3.1. Debiasing techniques 189
   5.3.2. How not to debias framing effects 191
   Expertise does not reduce framing effects 191
   Increasing reflection does not reduce framing effects 193
   5.3.3. Better debiasing strategies for framing effects 195
   Environmental debiasing techniques for framing effects 195
   (A) Dispelling cognitive illusions through information design 195
   (B) Rebiasing toward better decisions 199
   5.3.4. The power of environmental debiasing 201
   Why is environmental debiasing more effective against framing effects? 201
   Information design and rebiasing are quite different strategies 203

§5.4. Control over decision-making and the theory of action 206
   5.4.1. Anti-intellectualism and the phenomenon of the background 206
   Expanding the model of agency 206
   Background control 207
   5.4.2. The limits of background control 210
   Background control and framing effects 210
   Environmental debiasing is crucial in overcoming decision myopia 211
   Scaffolded autonomy 212
   5.4.3. Conclusion: Indirect intellectualism and its implications 213
   Indirect intellectualism 213
   Practical implications 214
   (A) An externalist approach to agentive control 214
   (B) Environmental debiasing as a public policy issue 215
   Coda: Environmental debiasing and the problem of manipulation 216
   (C) Debiasing and libertarian paternalism 216
   (D) Is environmental debiasing manipulative? 218

Conclusion 220
   Control: Its kinds and its limitations 220
   Indirect intellectualism, old and new 221

Bibliography 223
   Ancient works 223
   Modern works 224
## List of figures

The Watching Eyes Effect (from Bateson et al. 2006). xii

1.1: The structure of animal agency. 10
1.2: The intellectualist interpretation of the structure of human agency. 14
1.3: The circular structure of human agency. 55
1.4: The complete characterization of human agency. 73

2.1: A theoretical landscape of intentional action. 78
2.2: The two main perspectives on intentional action. 89
2.3: The T-800 smiles / The T-1000 smiles. 97
2.4: A (more complete) theoretical landscape of intentional action. 104
2.5: Seven records of one subject’s eye movements (Yarbus 1967). 109

5.1: The Ebbinghaus illusion. 178
5.2: The Ebbinghaus illusion, plus some environmental scaffolding. 179
5.3: The value function proposed by Tversky & Kahneman (1981). 182
5.4: Hyperbolic discount curves (taken from Ainslie 2005). 189
5.5: The original payday loan envelope. 201
5.6: A ‘debiased’ payday loan envelope. 202
5.7: Two visual aids found effective in debiasing framing effects (Garcia-Retamero & Galesic 2010). 203
5.8: Image of eyes looking after some bicycles (Nettle et al. 2012). 204
5.9: An intellectualist model of human agency. 211
5.10: Adding decision-making to the human agency model. 212
5.11: Adding the background to the model of human agency. 215
Introduction
In a way, everything that follows is an attempt to unpack what this image, and the story behind it, has to teach about the kind of agents we are, to what extent we are in control of what we do, and what we can do to increase and protect our autonomy.

The story takes place in a coffee room that works on the basis of an ‘honesty box’ mechanism: users agree to deposit the cost of whatever they consume in a box. A note next to the box gives them information about the prices of coffee, tea, and milk. No one supervises users’ behaviour. Hence the ‘honesty’ part.

Researcher Melissa Bateson and colleagues (2006) decided to adorn the price note with images that alternated on a weekly basis. First it was an image of a pair of eyes looking straight at the viewer; the next, the eyes were gone, and now some flowers were in their place. The next week, some other eyes, and so on. By counting the users’ contributions to the honesty box, the researchers found that when the sign was accompanied with images of eyes, users’ contributions were 2.76 times higher than when the sign was adorned with flowers. So it seems that something as irrelevant as a picture can dramatically influence people’s behaviour—and not just any behaviour, but intentional, calculative behaviour, of the kind we would have a hard time denying the title ‘intentional action’. Because, think about it: each coffee room user was supposed to notice the price of each consumed item, add it all up, and deposit the total amount in the box, which seems to demand the use of attentional and executive control. Moreover, this is called an honesty box, so people are expected to act morally, even if no one is watching. So the study suggests an entirely unrelated image can noticeably alter reflective action.

But one may justifiably think the behaviour was not all that reflective. After all, users go in and out of that coffee room every day, and use these products and the box all the time, so they can’t be expected to pay attention to the sign, calculate the precise costs of the stuff they drank, and put the exact amount every time. It is more likely that they use a quick heuristic to intuitively come up with a price that feels about right, toss something close to that value in the box, and be on their way. And if they do not have change with them, they may think “I’ll put twice as much in tomorrow”. Seen from this angle, the process is probably more automatic and habitual than reflective. But if the process is so over-practiced, why does a new, unrelated image bear such a great influence on it? Shouldn’t the activity be so automatic that the image would go by unnoticed and have no impact on the amount people put in?

Reflective, automatic, or a bit of both, the fact is that eyes led people to give more than twice as much the amount that flowers did. This raises some worrisome questions about human agency, since it seems to go against an image of ourselves that is very dear to us (by ‘us’ I mean
‘many people in Western, democratic, late-capitalist societies’): the image that we are in control of what we do.

One may try to fend off the worries by thinking this must be a rather isolated incident, due in large part to the study’s particularities. But that is not the case. In fact, Bateson et al.’s findings have been replicated by other studies. The Watching Eyes Effect, as it is sometimes called, is much more pervasive than one may initially think. And the Effect, in turn, is just one case, taken almost at random, of a much broader body of evidence amassed by many independent research programmes in the cognitive sciences. We have evidence for automatic processes affecting our reasoning and evaluation, informing our judgments our decisions, and influencing multiple aspects of social interaction (see “A typology of everyday automaticity” below for a sample of the literature). These studies make it an almost inescapable conclusion that a great deal of our behaviour is influenced by automatic cognitive processes of which we are often unaware, that are often unresponsive to our occurrent intentions, and that are therefore very difficult, if not impossible, for us to intentionally control.

A typology of everyday automaticity

The following typology does not seek to be exhaustive, but merely representative of the main kinds of automatic influence on behaviour documented in the empirical literature. I divide it in a cognitive and a social section, mostly because they have been studied separately in cognitive and social psychology.

[I] Automaticity in human cognition

- Automatic biases in reasoning — A strong body of research has documented multiple ways in which automatic processes bias our assessments of argument validity, probability and frequencies, among other kinds of reasoning. In one study (Evans et al. 1983), researchers asked participants to consider the validity of some arguments while assuming that all the information given was true. Is this argument valid?

  No addictive things are inexpensive.
  Some cigarettes are inexpensive.
  Therefore, some addictive things are not cigarettes.

  And this one?

  No millionaires are hard workers.

---

1 The mere appearance of being watched, for instance, has been found to generate increases in cooperative behaviour when playing dictator games and public goods games (Haley & Fessler 2005, Burnham & Hare 2007). The effect holds even if the appearance is generated by just three dots in a face–resembling configuration (·,:) (Rigdon et al. 2009).

Some rich people are hard workers.
Therefore, some millionaires are not rich people.

The former argument was deemed valid by 71% of the participants, and the latter by only 10% of them, despite both arguments having the same logical structure. Researchers argue that this pattern of response is explained by belief bias: the first argument’s conclusion is believable, and this automatically generates an intuitive feel of correctness, against which reflection has to compete in order for the participant to recognize the argument as invalid (which it is). But the second argument’s conclusion is unbelievable, and this generates a feeling of incorrectness. The response pattern suggests that most participants answer on the basis of intuition rather than reflective examination. Thus, automatic belief-based intuitions interfere with, and threaten to undermine, reflective control over our reasoning. 3

A different, quite famous example revolves around the following description:

Linda is 31 years old, single, outspoken, and very bright. At University, she studied philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. (Tversky & Kahneman 1983)

Think about how likely it is that Linda is:

1. A bank teller
2. A feminist
3. A feminist bank teller

People tend to think that 3 is more likely than 1, but this breaks one of probability's basic rules, the conjunction rule (i.e. the probability of a conjunction must be smaller than the probability of each of its components). Although it is impossible, the idea that 2 and 1 together are more likely than 1 enjoys an intuitive feeling of rightness due to its match with the description.

A similar phenomenon occurred when high school students were asked to estimate in a very short time the result of either

\[ 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 \]

or

\[ 1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8. \]

Those estimating the former sum provided a much greater result (2,250 on average) than those estimating the latter (512 on average) (Tversky & Kahneman 1974). This is one among a myriad studies that demonstrate the phenomenon of anchoring: when assessing an unknown quantity, we tend to base the estimate on an available starting point—even if the starting point is totally irrelevant or inadequate for the operation—, and the effects of this on our estimation may remain inaccessible to consciousness. In the case mentioned here, people seem to have anchored their estimate on the basis of the first few numbers in the sequence.

- **Framing effects on judgment and decision-making** — Automaticity plays a remarkable role not only in our reasoning, but also in everyday judgment and decision-making. Particularly, when choosing

---

3 “What is most interesting about belief bias experiments is that they show the limited control that we, conscious persons, have over our own thinking and reasoning. ‘We’ really reside in our reflective minds and enjoy an illusion of being in control of our own behaviour” (Evans 2010b, 117).
between two alternatives, the choice people make can turn on its head depending on how the alternatives are framed. In one study (Shafir 1993), participants were asked to imagine they were judges in a child custody case. They were given the following information:

Parent A has average income, average health, average working hours, reasonable rapport with the child, and a relatively stable social life. Parent B has above-average income, a very close relationship with the child, an extremely active social life, does a lot of work-related travel and has minor health problems.

A group was asked who should be awarded custody; 64% chose B. Another group was asked who should be denied custody; 55% chose B. Although the options in each case were ultimately equivalent, framing the question in terms of choosing someone vs. in terms of rejecting someone seems to automatically make some features more salient than others. Because B has the most positive traits (greater income and closer relationship with the child) B gets picked in one condition; but since B has the most negative traits (a lot of social time and travelling) B also gets rejected in the other. Intuition seems to automatically give salience to some features rather than others depending on the task’s framing, thus constraining the information that reflective processes have to work on. Framing was originally shown to alter people’s judgments and decisions when the options are presented as losses vs. as gains in a context of risk; but it also influences decision-making in risk-less contexts, e.g. when people are asked to evaluate an object that is portrayed in positive vs. negative terms—as in the example above—, or when they are asked to perform different tasks, e.g. attaining a positive outcome vs. avoiding a negative one (Levin et al. 1998).

Another well-documented framing effect occurs when one of the alternatives is presented as the default; when this is the case, people tend to go with the status quo. A great example was found by researchers studying the rates of organ donation in countries within the European Union (Johnson & Goldstein 2003; cf. Abadie & Gay 2006). These countries are sharply divided between those that have very low rates of organ donors (e.g. 4.25% of the Danish population, and 12% of the German population) and those with very high rates (e.g. 85.9% in Sweden, and 99.98% in Austria). Researchers found that the explanation for these differences was unrelated to issues like culture, political inclination, or language. It all boiled down to how the options were presented to people: in the former group of countries, no one was an organ donor until they registered to be one; in the latter, everyone was an organ donor until they opted out. Deviating from the default imposes a cognitive cost, so presenting an option as the default has automatic effects on human agents, who tend to solve problems with as little cognitive resource expenditure as possible.
[II] Automaticity in social behaviour

- **Automatic imitative behaviour** — Participants work on a task with a confederate who shakes his foot; the confederate is then replaced by one who rubs his nose. In each case, participants imitated the actions performed by the confederate without the confederate’s influence being reported by them (Bargh & Chartrand 1999). Researchers hypothesize this automatic imitative behaviour intensifies pro-social feelings and positive connections: people report liking a confederate who imitates their physical gestures more than a confederate who does not; furthermore, participants who are imitated by others cooperate more with them than with those who do not imitate them (Dijksterhuis et al. 2007, 71–73).

- **Automatic activation of social categories and stereotypes** — In one of the most widely discussed studies showing how the automatic activation of a social category affects behaviour, university professors (both male and female) tended to give higher evaluations to an undergraduate student’s CV if it had a male name than if it had a female name, all other things being equal (Moss-Racusin et al. 2012); on average, the professors (both male and female) evaluated the male candidate as more competent, and were willing to offer a higher entry salary and more mentorship opportunities to him than to her. Since the only difference between ‘him’ and ‘her’ was their name, it is inferred that the name automatically activated the associated ‘male’ or ‘female’ categories, including evaluative categories, all of which implicitly affected the professors’ evaluation process.

- **Automatic goal activation** — Not only do we automatically imitate others, respond to others’ imitations, and behave in accordance with implicit associations between social and evaluative categories: automaticity also modulates our motivational processes. Other studies suggest that people primed with the ‘college professor’ category do better than a control group in playing Trivial Pursuit and acquiring new knowledge (Dijksterhuis & van Knippenberg 1998), and that people primed with the ‘Apple’ brand are significantly more successful in creativity tests than those primed with the ‘IBM’ brand (Fitzsimmons et al. 2008).

[III] Automaticity, reflection, and intentional action

- **Confabulation** — If automaticity has such marked effects on cognition, evaluation, judgment, and decision making, then how are we able to account for our actions by citing reasons and intentions?

---

4 In this subsection I illustrate everyday social automaticity by mentioning social-psychological studies, many of which use priming methodologies to obtain the observed effects. I should mention that several priming-based results have been recently put into question due to recent reports of failure to replicate a number of famous studies in the field. The replication debate is not settled yet; see Doris (2015, 44–49) for a detailed account. I agree with Doris in that, though there will surely be several well-regarded studies and practices that will be dismissed due to the “RepliGate” scandal, it would be unwarranted to conclude from this that the entire domains of priming studies or social psychology will go down with them. As a philosopher engaging with empirical literature, one should try to be cautious and attentive to worrying failed-replication signs, and avoid leaning too heavily on only one study or group of studies. That said, the best evidence currently available, as revealed by relevant meta-analyses (DeCoster & Claypool 2004, Cameron et al. 2012), suggests that the automaticity phenomena discussed in what follows are robust overall, even if particular studies may not be.

5 There is a debate about how to philosophically interpret this experiment, and what its philosophical upshots are about the structure of our decision-making systems and the power of reflection over automatic processes. I return to this below (Chapter 3).
A significant amount of research suggests that these reflective accounts are often no more than **post-hoc confabulations**. The classical study along this line comes from Nisbett and Wilson’s (1977) seminal work on introspection. In one of their experiments, they asked mall visitors to pick the best-quality item from four sets of nylon stockings lying on a table in front of them. After choosing, they were asked to explain why they picked the stockings they did. What people did not realize was that all four sets of stockings were identical. Since researchers found a strong tendency towards choosing the stockings further to the right, the cause of their choice must have had to do with the simple fact of positioning. But no participant adduced that as a factor playing a role in their choice, and virtually all of them denied any possible influence of positioning on their decision when explicitly asked about it. Of course, positioning information was certainly perceptually available to them, and easily accessible for report; but the impact that it had on their choice was not. In cases like these, people’s explanations turn out to be confabulations, i.e. interpretations that seem to rationalize their behaviour.

There are questions about the solidity of Nisbett and Wilson’s conclusions (see Newell & Shanks 2014 for recent discussion), but introspection’s limitations when it comes to identifying the causes of behaviour—and our subsequent tendency to confabulate plausible but fake accounts—are backed by a great amount of evidence (Wilson 2002, Carruthers 2011).

• **The role of automaticity in intentional behaviour production** — By now it may seem clear that, according to empirical findings, automaticity has a pervasive influence on everyday behaviour. But philosophical argument provides reasons to think that the threat of automaticity runs even deeper, not only through aspects we do not seem to control, but also through the behaviour we apparently do. This is because even what seem to be intentionally controlled actions constantly rely on automatic processes (cf. Strawson 2003, Wu 2013a). Think about the complex, and by and large automatic, cognitive or sensorimotor processes we deploy when recognizing a face or a tone of sarcasm, when constructing a sentence or deciding what to eat at a restaurant. Even the simplest cases of mental action (like memory retrieval or imagination) and bodily action (like walking or retrieving a key from one’s pocket) call upon a broad number of automatic-associative routines that seem to be outside of reflective control—often outside of conscious access—, but whose operation is nonetheless necessary for our intended action to occur. To what extent are our intentional actions just higher-order ballistics, fully constituted by clusters of lower-order ballistics?

All of which raises the question that will be this work’s central concern: What kind of agency and control do we humans really have over our behaviour, if most of it is automatic? How does it work, and what are its limits?

This question brings serious concerns for political philosophy. People have long worried about the autonomy-undermining influence of things like advertising and ideological structures, and some sensible interpretations of the research seem to provide solid grounds for such worries. In his review of the science of human reasoning and decision-making, Keith Stanovich expresses
concern about the implications that the limitations of rational decision-making may have for individual autonomy:

Being rational means being unaffected by contextual factors that we deem irrelevant to the decision. If these contextual factors come to affect our decision, then it means that we lose personal autonomy to those who control those very factors. We give up our thinking to those who manipulate our environments, and we let our actions be determined by those who can create the stimuli that best trigger our shallow automatic processing tendencies. We make the direction of our lives vulnerable to deflection by others who control our symbolic environment—to those who control which frame we are presented with; to those who control what is in the choice set and in what order the alternatives are presented to us; and to those who control what we should consider the status quo. (Stanovich 2010, 46-47)

This makes it more urgent to wonder about the nature and limits of our agentive control, as well as about what we can do to protect it.

* * *

An attempt to answer these questions must necessarily bring different disciplines into dialogue. Questions about agency and control are often discussed theoretically by philosophers, and empirically by psychologists and other cognitive scientists. This work stands at the intersection of the two realms, in an exercise of what can be called ‘naturalistic philosophy’. I am sympathetic to the view that philosophical discussion should be sensitive and responsive to the empirical observations of the sciences, and engage in interdisciplinary dialogue whenever it may be fruitful.6

Naturalism is sometimes conceived as a novel approach to philosophy, when opposed to the armchair methods used e.g. in Descartes’ Meditations. But that is not correct, because most of the founders of Western philosophy were naturalistic philosophers.7 The so-called Presocratics were often called ‘naturalists’ (phusikoi) because their philosophical reflection was seamlessly intertwined with whatever empirical understanding of astronomy or other fields was

---

6 For diverse naturalistic approaches along these lines, see e.g. Habermas (1990), Railton (2004, 271), Doris (2015, 12–14). The naturalistic approach to philosophy—roughly, the view that philosophy should be sensitive and responsive to the empirical observations of the sciences—does not imply rejecting armchair methodologies, but rather rejecting the view that philosophy can work only on the basis of armchair methodologies (Levin 2013).

7 Socrates and Heraclitus were, perhaps, exceptions to this trend.
available at the time. Platonic dialogues like the *Timaeus* reveal thorough engagement with scientific theories, and Plato’s Academy was known as a place where scientists from all corners of the Greek world would gather to discuss their theories concerning geometry, astronomy, and other scientific fields. And Aristotle pioneered scientific investigation into living beings and, most noticeably, animals. So philosophical naturalism is as old as philosophy itself. If anything, armchair philosophy is the new trend. The recent movement toward empirically-informed philosophical reflection should thus be seen as a return to philosophy’s roots.

Among the ancient philosophical naturalists, Aristotle is particularly relevant to this project, since he was arguably the first Western philosopher to produce a systematic theory of action. (His theory of action was, of course, influenced by Plato’s reflections on the subject, but the dialogues, for all their relevance, do not yet contain a systematic account.) This theory stems from his biological investigations, and reconstructing it requires scanning the continuum that goes from his research on animal cognition and motion, through his treatise *On the Soul*, to his ethical and political texts. One of this work’s goals is to put Aristotle’s discussions of human agency into dialogue with the current debates, in search for mutual illumination. I think that contemporary research on habit, automaticity, and control can benefit greatly from Aristotle’s treatment of those very topics, and, vice-versa, an engagement with contemporary empirical literature can shed light on interesting aspects of his texts that may otherwise go unnoticed.

So this is a work of bridge-building. On the one hand, it attempts to contribute to the construction of bridges between the philosophical problems surrounding action and the empirical discussion of automaticity and control. And on the other, it seeks to contribute to building bridges between philosophy and its history. I embark on these two projects with the hope that the body of work at one side of each bridge can illuminate that at the other.

* * *

A natural place to start is at the origin of the discussion. I believe Aristotle entertained a question fairly similar to this work’s central concern. In fact, he thought that crucial aspects of human action were not determined reflectively, by reasoning: our dispositions to perceptually and semi-perceptually experience things as attractive, pleasant and good (or as repellent, painful, and bad) is determined by the non-rational cognitive processes that are shaped by habituation and have been sedimented into our character. Reason in this respect is secondary: it mediates between the character-determined goals, and the particular actions that we produce. But he also
argues that it is primary: reason should govern over passion and desire. But how can reason do this, if our primary goals cannot be intellectually determined? Aristotle’s answer, and the question itself, has remarkable resonances with a certain approach to our contemporary problem. In Chapter 1, I contend that part of his answer is that in order for us to take control over our life’s goals we must carefully shape our habits. But, the first steps of the process of careful habituation largely depend on other people: our parents and families, our friends, our teachers, our legislators… So this ‘care of the self’ turns out to be a political process. And reason plays its crucial role here: not as the calculative faculty of an individual agent, but as the social and political environment of institutional organization that sets up the conditions of possibility for individual rationality. The Chapter’s final upshot is that from Aristotle’s texts emerges an often-hidden sense of ‘logos’ as a shared common resource that requires care in order to allow for individual logos (and therefore individual agentive control) to flourish.

The latter claim, that individual control over actions is strongly reliant on the background of properly structured social practices, spaces, and institutions, is the core claim I want to ground in the following chapters through a discussion of the contemporary literature. It is a lesson latent in the Watching Eyes Effect image: rational capacities inside the head are often not sufficient for agentive control; environmental and public structures that empower our cognitive capacities are also necessary.8

In Chapter 2 I describe the discussion’s broad theoretical context. Two ways of conceiving the ‘intentional’ in ‘intentional action’ have been engaged in a long and arduous theoretical dispute. Intellectualists, on the one hand, consider that intentional action requires the use of reflection to coordinate behaviour in the service of an intention. Anti-intellectualists, on the other, claim that intentional actions can be produced without any top-down coordination with a view to an intention. Examples of this may include the skilful ways in which living beings navigate their environments, the many habitual and highly fine-tuned activities that we perform every day, or the actions of highly-skilled, expert performers. I describe how the growing realizations that, first, human action is thoroughly embodied and deeply embedded in its context, and second, human reflective capacities are severely limited, have rendered classical forms of intellectualism increasingly implausible, but have also spawned new forms of more situated intellectualism, which are more able to incorporate anti-intellectualist insights and vocabularies (about embodiedness, situatedness, and embeddedness) within its own framework.

8 This is, of course, not a novel view. It is a variation on a theme inspired, among others, by Merleau-Ponty’s phenomenology, notably developed by the extended-mind approach (see e.g. Clark 1997; Clark & Chalmers 1998; Heath 2014), and also explored by a recent trend in public policy (Thaler & Sunstein 2008).
The current debate is between anti-intellectualists and situated intellectualists, and revolves around how best to explain the large portion of our behaviour that occurs on the basis of automatic cognitive processes.

The first step towards advancing the debate is therefore clarifying what we mean by 'automatic'. This is a very messy concept, both in the theoretical and empirical literatures. In Chapter 3, I present an overview of the dual-process literature’s recent history (its ancient history goes back at least to Plato and Aristotle), in order to precisely define the concepts of intuition and reflection. These definitions turn on the psychological concept of 'working memory', a construct that is very useful in understanding how human rationality works, and how it does not. With these tools in hand, I assess several common conceptions of automatic processes, and argue they should be rejected. Instead of conceiving automatic processes as mechanical, ballistic, largely unconscious, and therefore uncontrollable, evidence suggests that automatic processes can, and often do, produce pre-reflective, fine-grained, and norm-sensitive behaviours. The normativity of these behaviours is based on the completion of environmental patterns (a property that has been recently described as the alleviation of a “felt-tension”, or a “directed discontent”, and is very similar to Merleau-Ponty’s notion of motor intentionality). If automaticity has a normativity of its own, then there is a real chance that there are full-blown reflection-independent intentional actions, as anti-intellectualists have argued.

So, are there reflection-independent, fully automatic intentional actions? To move toward an answer, I turn in Chapter 4 to what anti-intellectualists have often considered to be their mainstay cases, and the territory less hospitable to intellectualist strategies: skilled action. Recent versions of situated intellectualism try to show that some reflection is necessary even for the production of skilful actions, because automaticity, while norm-sensitive and necessary, is still insufficient to account for it. This would be very bad news for anti-intellectualism. In their defence, anti-intellectualists have recently highlighted three psychological phenomena that seem to resist any intellectualist explanation: expertise-induced amnesia (experts seem unable to explain how or why they do what they do while performing), choking under pressure (reflecting on what one is skilfully doing while doing it seems to hinder rather than aid performance), and expert confabulation (when experts do get to report the reasons or rules underlying their behaviour, they tend to misdescribe them, which suggests reflection serves only to produce post-hoc rationalizations). Anti-intellectualists argue that these phenomena show how reflecting while skilfully performing is at best unnecessary, and at worst harmful, to skilled action. But I argue that, arguments to the contrary notwithstanding, intellectualists have the upper hand in this
debate: they can produce sensible explanations for all three phenomena, and can build a strong case for the need to reflect even while performing habitual, everyday actions.

Does this mean the victory of intellectualism? Not quite. The anti-intellectualist has always had at her disposal two main strategies to prove her point, and arguing that reflection-free actions exist is but one of them. The other one is arguing, like Aristotle, that reflection is significantly constrained and informed by a background of more fundamental, non-reflective cognitive processes, and therefore, that whatever level of reflective control we have turns out to depend on how well the background automatic and intuitive processes frame reflection’s activity. Recent anti-intellectualism has tended to relegate the latter strategy to second place, and focus on the former. Since Chapter 4 shows the former to be rather unsuccessful, in Chapter 5 I turn to the other one. I argue that the pervasiveness of framing effects on reflection demonstrates the insufficiency of reflective control to provide us with a satisfactory level of agentive control, since reflection can be detoured when the practical environment makes irrelevant features intuitively salient, or diminishes the salience of highly relevant features. I describe several cases that reveal this feature of our agency is widely exploited by actors interested in manipulating our limited reflective decision-making capacities. This leads to the conclusion that individual reflection warrants agentive control only to the extent that it is supported by a well-designed environment, i.e. an environment that bolsters, rather than undermines, the cognitive powers of individuals.

This view implies a return to Aristotle’s conception of reason as a common public resource, necessary for the proper functioning of individual reflective capacities, and dependent on smart environmental design directed towards bolstering individuals’ reflective capacities. Individual agentive control turns out to depend on a fertile soil of public care.
Chapter 1: Habit and reason in Aristotle’s theory of agency

This work explores the roles that habit and automaticity, on the one hand, and reason and reflection on the other, play in shaping human control over action. I called the central question of control: ‘What kind of agency and control do we humans really have over our behaviour, if most of it is automatic? How does it work, and what are its limits?’ As this Chapter makes clear, it is an old question that received much attention from the first Western thinker to offer a systematic theory of action. I argue that Aristotle’s treatment of the roles of habit and reason in action production can serve as a heuristic tool for contemporary discussions of the issue. This Chapter engages with Aristotle’s texts on human (and animal) action, and some interpretative debates surrounding it.

Let us begin with a paradox: Aristotle holds both that animals are agents and that they are not. In some contexts he treats the concept of ‘action’ as extensible to all animals and children, and in others he treats it as proper to adult humans. To interpret this consistently, we must assume that there are two senses of agency at play: a broad one in which both animals and humans count as agents, and a narrower one in which only human adults do. Given this, in order to clarify the structure of human agency it makes sense to start by asking: what is it about human adults that makes us agents in a more restrictive sense than animals or children?

A somewhat obvious path toward the answer seems to be to say that we have something that animals and children lack, and that this exclusive trait gives us greater agentive control. Aristotle does hint in that direction, by claiming that we are the only animal that has logos (e.g. Politics [Pol.] I.2 1253a9–18), and also that we can attribute action to adult humans alone because they are the only ones who act by means of reasoning (logismos) (EE II.8 1224a28-30). Thus it seems straightforward that for Aristotle the distinguishing feature of human agency is the rational faculty, and the things it allows humans to do or achieve, among which deliberation (boulē), decision (prohairesis), and the virtue of practical wisdom (phronēsis) are particularly relevant.

Against this intellectualist approach, I argue below that for Aristotle the key structural difference between animal and human agency does not reside merely in our possession of the rational faculty. The story is much more complicated, and it crucially involves something that other animals have and we lack. For nature provides them with a certain practical guidance that we are largely deprived of, and must therefore compensate for: whereas non-human animals have

---

1 In both the Eudemian Ethics [EE] (II.6 & 8) and the central books (common to both Ethics [EN]; particularly EN VI.2) Aristotle deprives animals and children of agency. But he grants it in EN III.1 & 2, and at least implicitly throughout On the Motion of Animals [MA] and in On the Soul [DA] III.9–11 (see also Physics [Phys.] II.8 199a20–ff).
natural dispositions that largely determine which things appear good or bad to them in perceptual experience, we instead must develop habits that determine what appears good and bad to us.\textsuperscript{2} My thesis is that, whereas practical reason makes us quantitatively more complex agents, what makes our structure of agency qualitatively different from that of other animals is the fact that we can shape the way things appear to us (our ‘appearance of the good’) by shaping our habits. So the mere possession of reason cannot account for properly human agency. Reason does play a crucial role, but not as it is usually conceived (a deliberative capacity that mediates between perception and motion): its crucial function is to carefully shape our habits. Now, since habit-care starts in each individual long before she can use her own reason, habit-care turns out to be a necessarily social endeavour. Because of this, I defend an intellectualism about human action that is indirect in two ways: because reason can determine our life’s goals only by shaping habits, and also because an individual’s ability to rationally shape her own character depends on her having received proper early care from others early in her life. The final upshot will be a view of reason, not as a faculty that resides in an individual’s soul, but rather as a common resource shared and nurtured by a community: a set of social and political structures that make individual rational agency possible.

After (§1.1) briefly reviewing Aristotle’s agency vocabulary and describing the principles of minimal (i.e. animal) agency, I discuss (§1.2) traditional intellectualist interpretations according to which the rational faculty is the crucial difference between animal and human agents. I (§1.2.1) present the main variations on the intellectualist theme and then (§1.2.2–1.2.3) argue that this account misses something crucial: the view that reason establishes the very principle of our action, i.e. the goal, goes against solid textual evidence. I therefore (§1.3) construct an alternative interpretation, largely based on crucial dialectic (\textit{EN} III.5) between Aristotle’s view and that of an unnamed opponent, arguing that, unlike any other animal, we humans can control our action’s goals by caring for and shaping our habits. I show then how (§1.4) Aristotle considers habit-care to be a task proper to the political community, and how reason plays a determinant role by means of the laws and the constitution. My conclusion is that for Aristotle properly human agency depends on the political practices of reason that make self-care possible. This interpretation has the advantages of retaining the original intellectualist intuition that reason is the crucial differentiating factor between animals and human adults, while doing justice to Aristotle’s claims that character, and not reasoning, determine the agent’s goals, and also highlighting the connection between rationality and that other distinguishing aspect of human animality: the fact that we are the most political of all animals. We are rational beings, but we can enact our rational nature only within the proper political context.

\textsuperscript{2} This may sound like animals have no habits and humans have no natural dispositions, but that would be incorrect, as I explain in more detail later.
§1.1 Principles of animal agency

1.1.1. Aristotle’s vocabulary of action

What does Aristotle mean by ‘action’? First of all, in the Aristotelian texts an action is a kind of locomotion. The canonical discussions of animal movement (DA III.7–11; MA.6–11) treat the verbs ‘to act’ (prattein) and ‘to move’ (kineisthai) largely interchangeably, and the word ‘action’ (praxis) appears in connection with an animal’s locomotive pursuit and avoidance of some object (DA III.7.431b8–10). Aristotle calls locomotion ‘voluntary’ and labels it ‘action’ whenever the origin or principle of motion is internal to the mover (EN III.1.1111a22–23; EE II.6).

Thus, the notion of an origin or principle of motion is crucial. It gets expressed in several different ways: Aristotle claims, for instance, that agents are kurioi of their actions, and that our actions are eph’ hēmin. The word ‘kurios’ comes from the political context, meaning ‘master’ or ‘lord’; Aristotle gives its significance an abstract turn as ‘commander’ or ‘controller’. I am basically translating the word when I claim Aristotle says “we are in control of our actions”. On the other hand, the expression ‘eph’ hēmin’ became the standard technical notion in later Greek philosophy to discuss problems related to action and responsibility after Aristotle introduced it.

To use Aristotle’s expression, saying that X is eph’ hēmin, or up to us, means that it is up to us to do it or not to do it, i.e. that whether X happens or not depends on our causal influence in the world. All these notions are linked together in a riddle-like passage from the Eudemian Ethics:

\\(\text{ ámbte ὅσων πράξεων ὁ ἀνθρωπός ἐστιν ἀρχὴ καὶ κύριος, φανερῶν ὅτι ἐνδέχεται καὶ γίνεσθαι καὶ μὴ, καὶ ὅτι ἐφ’ αὐτῷ ἐστιν ἢ μὴ ἐστιν, ἢ μὴ ποιεῖν, αἰτίος τοῦ ἐπιτῶν αὐτὸς ἐστιν καὶ ὅσων αἰτίος, ἐφ’ αὐτῷ.}\\)

So it is clear that for all actions of which man is a principle and of which he is in control [kurios], it is possible for them to come to be and to not come to be, and that, of those actions of whose being or not being he is in control, it is up to him [eph’ hautoi] whether

---

3 E.g. MA 2.698b21-26, 7.701a7-9.

4 See e.g. his version of this claim in EN III.5: “ei δὲ […] μὴ ἔχομεν εἴς ἄλλας ἀρχὰς ἀναγαγεῖν παρὰ τὰς ἐν ἑμῖν, ὡς καὶ αἱ ἀρχαὶ ἐν ἑμῖν, καὶ αὐτὰ ἐφ’ ἑμῖν καὶ ἐνδόθησα. [If […] we cannot refer back to other origins besides those within us, those things whose origins are within us are also up to us and voluntary.]” (1113b19–21)
they come to be or not. And he is the cause \[aitios\] of those things that it is up to him to do or not do; and those things, of which he is the cause, are up to him. (EE II.6 1223a4–9)\(^5\)

Thus agency is based on the ability to produce bodily movements that stem from inner principles of motion. Such motions can hence be called ‘actions’, and insofar as they are generated by inner principles, they are voluntary and up to us. Thus, given that the notion of action is based on the notion of inner principles of motion, the difference between human and animal agency should become clear by clarifying the principles of motion proper to each.

1.1.2. The principles of animal agency

What are the inner principles of animal motion? In a crucial passage, Aristotle distinguishes two:

\[όρῶμεν \text{ δὲ τὰ κινοῦντα τὸ ζῷον διάνοιαν καὶ φαντασίαν καὶ προαίρεσιν καὶ βούλησιν καὶ ἐπιθυμίαν.}\]
\[ταῦτα \text{ δὲ πάντα ἀνάγεται εἰς νοῦν καὶ ὄρεξιν.}\]
\[καὶ γὰρ ἡ φαντασία καὶ ἡ αἴσθησις ἐν ᾧ ἄλλοις διαφοράς. \text{βούλησις δὲ καὶ θυμὸς καὶ ἐπιθυμία πάντα ὄρεξις, ἢ δὲ προαίρεσις κοινὸν διανοίας καὶ ὄρεξεως· ὥστε κινεῖ πρῶτον τὸ ὄρεκτὸν καὶ τὸ διανοητὸν.}\]

We see that the things that move the animal are cognition, \textit{phantasia}, decision, wish, and appetite. But all of these come down to cognition [\textit{noûs}] and desire [\textit{orexis}]. For \textit{phantasia} and perception take the same place as cognition, since they are all discerning faculties [\textit{kritika}], although they differ from one another in the ways we have mentioned elsewhere. On the other hand, wish, spiritedness, and appetite are all desire, and decision is common to both cognition and desire; hence the things that move primarily are the object of desire and the object of cognition. (MA 5 700b17–24)

Here the faculties of perception, \textit{phantasia}, and intellect are all grouped together under the label ‘discerning faculties’ (\textit{kritika}) because they are all cognitive in a broad sense of the term: their function

\(^5\) At least since Alexander of Aphrodisias’ \textit{De fato}, scholars have discussed at length whether Aristotle was an indeterminist or a compatibilist, whether he endorsed some version of the principle of alternate possibilities, and in general what his view is or implies regarding the free will–determinism debate (for discussion and extensive bibliography, see Gauthier & Jolif (1970, 217–220), Taylor (2006, 166), and Destrée (2011)). Following this work’s general approach, I will bypass the free will–determination debate as much as possible, in order to focus on the problem of control, in the hope that this turns out to be a more tractable, and ultimately more fruitful, approach. Particularly with respect to Aristotle’s works, given that it is unclear that he directly dealt with the free-will problem anywhere, and since he did not use concepts clearly analogous to those like ‘will’, ‘freedom’ and ‘causal determinism’, I take it that we are better off if we avoid forcing such notions and problems into the interpretation: otherwise we may end up distorting rather than clarifying the actual texts. So I will rather focus on Aristotle’s way of dealing with the issue he explicitly cared about, namely why it is that we can consider ourselves to be in control (\textit{kurioi}) of our actions.
is to allow animals to discern (krinein) the world’s entities, and in that sense all three of them are producers of knowledge. So for all animals a cognitive, discerning faculty is one of the principles of action. The same goes for desire: animals can have desires for simple, immediate objects (i.e. ‘appetites’ or epithumiai), or more complex desires (i.e. ‘wishes’ or boulēseis, desires implying reason), but in any case some kind of volitive process must play a role in explaining why they move. Hence cognition and desire are within the principles of action, the inner sources of animal motion.

Now, these two are not independent capacities; cognition and desire are intrinsically linked, and a crucial link between them is pleasure: all animals who have perception have desire, Aristotle holds, because perception implies the ability to discern pleasure from pain, pleasant from painful objects; and if an animal perceives these, then it has the most basic form of desire: appetite (DA II.3 414b1–6). So the experiences of pleasure and pain are perceptual experiences, and vice-versa, perceptual experiences can be of objects as pleasant or painful. Therefore, perceptions (and the phantasiai that follow them) are motivational, in the sense that they direct the animal towards certain objects or away from others by presenting them as worthy of pursuit or avoidance, through pleasure and pain. This is visible in Aristotle’s account of why all animals must have some sense of taste:

\[ \text{ἡ δὲ γεύσις διὰ τὴν τροφὴν. τὸ γὰρ ἴδιον διακρίνει καὶ τὸ λυπηρὸν αὐτῇ περὶ τὴν τροφὴν, ὥστε τὸ μὲν φεύγειν τὸ δὲ διώκειν [...].} \]

Taste [belongs to all animals] because of nutrition: for by means of it they discern the pleasant and the painful concerning nutrition, so that they flee from the latter and pursue the former [...]. (De sensu 1 436b15–17)

---

6 The notion of ‘phantasia’ presents many interpretative complexities. I will rely here on the widely shared view that one of the things it does is allow animals (humans included) to experience previously perceived objects while they are not perceptually present to them. Phantasia therefore entails some kind of perceptual memory, and some level of mental-modelling capacities that are useful, e.g., in imagining the attainment of a goal that is not yet achieved. For more detailed discussion see Nussbaum (1978, essay 5), Frede (1992), Labarrière (1997), Schofield (2011), Moss (2012, Chapter 3), and Carbonell (2013).

7 I am using ‘cognition’ to translate all the cognate words noús, noēsis, and dianoia, which Aristotle tends to use interchangeably in action-theoretic contexts. However, as the previous passage makes clear, these words have at least two relevantly different meanings: the broader sense of ‘cognition’ includes all discerning capacities (reason, phantasia, and perception); but ‘cognition’ in the narrower sense is a particular discerning faculty different from perception and phantasia.

8 For a detailed account of the centrality of perception and pleasure for motivation and agency in general, see Moss (2012). This section follows her interpretative line.
Now, the fact that perceptual objects appear as pleasant or painful, i.e. as worth pursuing or avoiding, means that perceptual objects also appear as good or bad. In other words, goodness and badness, insofar as they coincide with pleasure and pain, are also perceptible properties of objects for an animal. ‘Perceiving something as good’ does not require some intellectual operation. It just means that in perception it is revealed as something worth being pursued; as to-be-sought, or to-be-done—that is, as pleasurable.

In the passage just quoted Aristotle talks about taste, but a similar analysis can be applied to the other sensory modalities: if an animal is to move, and find nutrition by moving, then it needs senses that guide it toward the pleasant things, and away from sources of pain (cf. De sensu 1 436b18–437a2). And pleasure and pain remain a feature of perceptual experience for all animals, from the simplest one to the most complex. The other cognitive capacities, phantasia and intellect, merely add complexity and widen the range of possible perceptual pleasures and pains; they elaborate on the perceptual pleasures and pains (by e.g. allowing animals to bring them to mind while absent, or re-describing them as ‘fine’ and ‘good’ rather than pleasurable), or allow the animal to compare pleasures and pains (thus allowing them to judge a small present pain as pleasurable because of the greater pleasure that it will bring forth later); but they do not provide a source of pleasure and pain independent of their perceptual origins (with the exception of the theoretical, disembodied pleasures of contemplation). As cognitive capacities grow more and more complex, pleasures and pains increase in complexity too. Hence, when an animal is able to recall the past, imagine the future, and attain experience, pleasure and pain can take the shape of emotions, or passions, like shame and pride, or fear and pity (Rhetoric [Rhet.] II, Pol. VIII.5 1340a12–14). Or pleasure can be experienced in the unimpeded performance of complex activities (ENT.8 1099a7–16; VII.13 1153b9–17). The key point is that, whether the object of desire is presently here, is remembered through past experience, or is conceived as an object of future experiences, all appearances of something as good or as pleasant trace back to perception as the main faculty through which we experience an object as pleasant or good.

Thus, pleasure links the faculties of cognition and desire in the animal, and links their activity toward the pursuit or avoidance of pleasant or painful objects. So the sensory, volitional, and motor capacities work together. This seems to be the meaning of Aristotle’s puzzling claim that perception, desire, and motion are the same in activity, but different in being (e.g. DA III.7 431a8–14). As Whiting (2002) argues, all these capacities, though different in definition, constitute a functional unity whose different aspects are always activated together, and which Aristotle sometimes calls “the faculty of desire [το ὠρεκτικόν]”, and we may follow Whiting in calling ‘the locomotive soul’—or ‘the sensorimotor system’ in contemporary parlance. So in animals
the capacities of cognition, volition, and motion constitute a unified faculty that, though complex, works in unison to generate animal action. Whenever this complex sensorimotor faculty causes the animal’s motions, we can say that the principle of motion is internal to the animal, and hence that the animal is in control of it.

One more element is missing from this account of minimal agency. For, although Aristotle claims that there is only one mover, namely the faculty of desire (understood in the complex sensorimotor way just described), he then goes on to say that the first mover is not said faculty but “the good itself”, i.e. the external object of desire:

εἴδει μὲν ἐν ἄν εἶη τὸ κυνόν, τὸ ὀρεκτικόν, ἢ ὀρεκτικόν, πρῶτος δὲ πάντων τὸ ὀρεκτόν (τοῦτο γάρ καὶ ὢν κανοφυμενον, τῷ νοηθῆναι ἢ φαντασθῆναι), ἀρθμῷ δὲ πλείῳ τὰ κανοίντα. […] ἐστὶ δὴ τὸ μὲν ἀκίνητον τὸ πρακτόν ἀγαθὸν, τὸ δὲ κινοῦν καὶ κινούμενον τὸ ὀρεκτικόν

The mover will be one in form: the faculty of desire, qua faculty of desire, but the first of all movers will be the object of desire [to orekton] (for it moves without being moved, by being cognized or being the object of phantasia), and the movers are many in number. […] So the unmoved mover is the good attainable in action, and the mover that is moved is the faculty of desire (DA III.10 433b10–13 … 15–17)

So the apparent good, the object of desire, seems to be the very first principle of action, prior even to the sensorimotor faculty’s activity, since it is its presence (sensed through perception or imagined through phantasia) that triggers the whole internal process. If this is so, the first principle of action seems to be external to the agent (Labarrière 1997). If this is so, no action would be produced by the agent’s internal principles, and animals would therefore not be agents. After all, everything that the sensorimotor faculty seems to do is react to the object’s presence.

However, even though the object is in a sense the first cause of the process of animal motion, still what moves the animal is not the object as such, but rather the object as pleasant or as painful. And the object is not pleasant or painful in itself, but only in an animal’s discerning experience of it. In fact, the object on its own is pleasant only potentially, and becomes pleasant in actuality only through the animal’s cognitive activity. So what makes a certain animal cognize a given object as pleasant or painful? This is determined by the animal’s natural dispositions, which are tailored to its nutritional and reproductive needs.
One half of [animal] life consists in the activities of procreation, and the other in the activities related to nutrition; for all their efforts and their way of life happens to revolve around these two. [...] And what is in accordance with nature is pleasant; and each animal pursues that which is pleasant in accordance with nature (History of Animals [HA] VIII.1 589a2–9)

So we should not identify the external object with the principle of motion. Rather, the principle of animal action is the animal’s natural dispositions of practical perception, which determine which objects become objects of desire by appearing pleasant (hence good and worth-being-pursued) or painful (hence bad and worth-running-away-from) for each animal in accordance with its own biological composition and needs.9

Thus the principles of minimal agency turn out to be two: the capacities of cognition and volition (which together constitute the locomotive or sensorimotor faculty), and the natural dispositions that determine what appears pleasant or painful in each animal’s discerning experience (see Fig. 1.1). Whenever the animal’s motion is explained by reference to these two principles, we can call such motion an action, and an animal an agent, in the broad sense of the words.10

---

9 Further evidence for this view is that the same perceptual object can be pleasant or not pleasant at different times, depending on the animal’s situation: “the smells [related to nutrition] are pleasant when we are hungry, but to those who are satiated and require nothing they are not pleasant” (De sensu 5 433b22–23). This shows that the crucial element is not the object, but the animal’s dispositions.

10 For contemporary accounts of minimal action that resonate with this proposal, see Merleau-Ponty’s notion of ‘motor intentionality’ and the accounts of sensorimotor agency inspired by it (Chapter 2, §2.1.2 below).
§1.2. Principles of human agency: intellectualism and its limits

If those two are the principles of animal agency, what then distinguishes human adult agency from it? In this section (1.2.1) I introduce the traditional view that identifies practical reason as the distinctive feature. Then (1.2.2) I object that this view has a crucial limitation, and (1.2.3) conclude by reviewing the way in which Aristotle himself replied to this objection.

1.2.1. Intellectualist interpretations of human agency

The extant corpus mentions many features that distinguish humans from all other animals. Some animals, for instance, have some sort of memory, but we are the only ones who can recall past events (HA I.1). Some animals can imitate, but humans are the most imitative of all, and learn their first lessons through imitation (Poetics [Poet.] I.4). Compared to other animals, humans have quite underdeveloped senses of sight, hearing, and smell, but their senses of taste and touch are the most accurate of all; and this in turn makes us the most intelligent [phronimotaton] among animals (DA II.9). Only humans take pleasure in smells that are not related to food, like those of flowers (De sensu 5 443b26–ff.). Man even has the smoothest skin of all animals because, in proportion to size, the human male emits more semen than any other animal (HA VII.2). The list goes on. But among these things that make us different from all other animals, what makes us more complex agents?

Many of the specific human traits point to the one that Aristotle is most renowned for: besides perception and phantasia, human cognitive capacities also include reason (logos) and calculation (logismos). After all, if we add up our capacity to recall past events, great imitative skills, and superior senses of taste and touch, these and other unique abilities add up to the claim that we are the animal that has reason. The general intellectualist outlook holds that the key distinctive feature of the structure of human agency is the possession and use of reason.

Here I am using the term ‘intellectualism’ in a different sense than usual. In the context of Aristotelian action-theory scholarship, it is more often understood as the view that reason determines the goals of human action, and is contrasted with the ‘anti-intellectualist’ view that character, and not reason, determines such goals. I will capitalize the word Intellectualism to refer to the goal-setting sense of the term; and use intellectualism in lowercase to refer to the humans-


12 See e.g. DA III.3 428a21–24, III.10; EN I.7 1097b33–1098a5; EE II.8; Pol. I 2 1253a9–10, VII 13 1332b5.
vs.-animals sense (i.e. the view that the crucial difference between humans and non-human animals is the former’s possession of reason). Although different, the two senses are not unrelated. In fact, Intellectualism offers a powerful argument for intellectualism: the Intellectualist view that humans rationally determine the goals they pursue can be used by the intellectualist to ground her view that rationality makes human agency more complex than that of other animals. Before discussing some Intellectualist arguments in the next subsection (§1.2.2), I want to distinguish different versions of the intellectualist claim. It can in fact be substantiated in different ways depending on the specific function that is assigned to practical reason.

Now, there are several interpretations concerning the role of practical reason in the production of action, but all of them must respond to two striking features of Aristotle’s psychology: first, the human soul is divided into non-rational and rational parts, the former associated with character and capacities we share with animals, the latter allowing for calculation- and deliberation-dependent activities and dispositions. And second, Aristotle claims that the irrational part determines the goals of action, whereas reason is in charge only of establishing “the things toward the goal”. Thus, the diverse kinds of intellectualism stem from different interpretations of the phrase “things toward the goal”, since they specify reason’s realm of operation in action production. I will now succinctly describe three of the most influential intellectualist interpretations present in the current debates.

*Instrumental intellectualism* — The most restricted interpretation of ‘things toward the goal’ takes it to refer simply to the *means* required for the goal. This gives rationality the purely instrumental function of identifying steps necessary to achieve an end, the latter being specified by the agent’s character. Such interpretation has been called ‘*Humean*’, since reason turns out to be little more than a ‘slave of the passions’, i.e. a mere assistant of non-rational goal-setting processes. However irrationalist this interpretation may seem, it counts as a version of intellectualism as long as it includes the claim that the distinguishing feature of human agency is the ability to assess the most efficient means, develop complex plans, and the like.

*Specification intellectualism* — Other interpretative possibilities open up if you consider the “things towards the goal” to include not just means, but also other things required by complex goal pursuit. A particularly influential line of interpretation holds that the goals that character specifies

---

13 See next note for references.

14 This is a view not often supported, but often mentioned as a relevant dialectical opponent (cf. e.g. Irwin (1975); McDowell (1998a); Price (2011); Moss (2012; 2014)). For a Humean reading of Aristotle, see e.g. Fortenbaugh (1964). Bear in mind the difference between intellectualism about the structure of human agency, and the more traditional Intellectualism about goal-setting. Anyone endorsing a Humean view of practical reason would not be able to endorse Intellectualism, but could still defend intellectualism.
are rather abstract (like ‘happiness’, ‘doing well’, ‘the fine’ or ‘the noble’), and therefore reason must get involved not only to establish the means, but also to determine what counts as attaining the goal in each practical situation—what counts as doing well, or as a fine and noble action, here and now, in this particular context. Practical reason’s function is thus not solely instrumental, but also specificatory: deliberation establishes means to achieve the goal, but must previously determine the goal’s specific content in each practical situation. An intellectualist could point to this function of practical reason as the distinctive trait of human agency.

Coordination intellectualism — Yet another possible intellectualist interpretation, arguably within the bounds of the Aristotelian division of labour between character and reason, is the view that reason is in charge of assessing whether the currently active goal is in accordance with broader relevant considerations. This may include assessing whether the goal fits with one’s overall conception of the good, whether it is preferable to competing goals, or whether seeking it would generate undesirable consequences. If reason determines that a given goal is not in accordance with one’s practical concerns, then it can override it, and decide on a different path of action. This interpretation gives reason a coordination role, which allows the agent to lead a consistent and unified life, responsive to the broader goals that character purportedly establishes, mainly by evaluating each apparent good, and rejecting or postponing it whenever necessary. Thus, “the eligibility of a target in context is tested by deliberation” (Price 2011, 152–153). This gives the intellectualist yet another function of reason whereby humans may be able to exert more control over their actions than other animals.

Although interpreters disagree about which of these interpretations of Aristotelian practical reason is correct, I need not choose between them here, for two reasons. The first is that they are not necessarily incompatible. In fact, a thorough intellectualist may stack all these accounts together and argue that human agency is structurally different from animal agency because of the greater control afforded by practical reason’s three functions. Reason receives the somewhat indefinite motivational input from the agent’s character, and structures it so as to make it specific, render it consistent with the agent’s broader goals and plans (or override it if
inconsistent), and provide a specific set of means whereby the goal is to be achieved. All of these features are unique to us, since they are different uses of calculation (*logismos*) and deliberation (*boulē*), capacities absent from all other animals. For them, the appearance of an object as good or bad immediately generates a desire, and desire leads immediately and invariably to action; but the human capacities for rational deliberation and calculation stand in between desire and action (see Fig. 1.2), making it possible for the agent to exert more control by allowing her to *not* act even in the presence of an apparent good. It is truly *up to them* whether the action occurs or not. Calculation and deliberation also allow human agency to operate on a much broader temporal structure: humans alone can coordinate their actions in the long term to develop complex plans. All of this supports the view that rationality is the key structural novelty that sets human agency apart from basic, animal agency. McDowell beautifully captures the foundations for the intellectualist intuition:

> The category of purposive behaviour, behaviour that can be explained by giving its end, extends to brutes as well as human beings. But human beings are special among animals in having a capacity for articulable thought. Purposive behaviour in brutes is an immediate response to an opportunity for gratification of non-rational motivational impulses […]. The peculiarly human capacity for thought allows for purposiveness without that immediacy; thought can mediate gaps between project and execution (1998, 23).

Figure 1.2: The intellectualist interpretation of the structure of human agency.

Besides the intellectualist accounts being compatible, the second reason for not choosing between them is that I will argue against the intellectualist outlook in general. I hold that even if all three of these proposals hold (i.e. even if all three functions of practical reason distinguish human and

---
17 As will become clear in Ch. 2–3, contemporary intellectualism attributes similar functions (planning, coordination, and inhibition) to reflective thinking. Modern intellectualism would argue that reflective thought is necessary for the production of any intentional action.
animal agents), human agency would certainly be more complex than the animal counterpart, but the claim that we are in control of our actions would remain unjustified. In what follows I argue that (§1.2.2) intellectualism is susceptible to an important objection; and that (§1.2.3) in a text in which Aristotle himself considers this objection, he gives a non-intellectualist reply.

1.2.2. The limitation of intellectualism: goal-setting

Recall the two peculiarities of Aristotle’s moral psychology mentioned above (i.e. the division between the soul’s rational and non-rational part, and the division of labour between them). Given these, any intellectualist account of human agency must face a crucial limitation, namely that the goals of action, which give the crucial input that reason processes and elaborates, are not under reason’s control. Given the character-reason division of labour, reason has no direct control over the appearance of the good (i.e. over which objects appear as desirable and which as avoidable): reasoning cannot make something appear good if character did not already present it as such. Something may, of course appear good by reasoning as a means to another end, or as a specification of an already given end, but this does not count because those rational appearances of the good, important though they may be, are dependent on a prior character-based appearance of the good. If this is so, then none of reason’s aforementioned functions provides us with any control over the principle of our actions.

That is, if reason does not determine the goals of action. But Intellectualism holds that it does. My view is that there are stronger reasons to reject Intellectualism than to endorse it, and therefore that intellectualism, as conceived thus far, fails to account for the kind of control that is proper to human agents. And Aristotle’s texts provide plenty of evidence against Intellectualism.

Goal setting is out of reason’s reach

I have mentioned that reason’s three functions are but applications of deliberative and calculative capacities, i.e. of *logismos*. Several lines of textual evidence, however, show that *logismos* cannot establish the goals of action. First, we cannot arrive at the starting points of practical reasoning through reasoning. Second, deliberation is not about goals, but only about things toward the goals. And third, goals are determined not by reasoning, but by experience, habit, and character.

(A) There is no reasoning of starting points

Consider Aristotle’s claim that there is no *logismos* of theoretical or practical starting points. Here is an argument from a passage that presents the division of labour between character and reason:
πότερον δ' ἡ ἀρετὴ ποιεῖ τὸν σκοπὸν ἢ τὰ πρὸς τὸν σκοπὸν; τιθέμεθα δὴ ὅτι τὸν σκοπὸν, διάτι ποιοῦν οὐκ ἔστι συλλογισμὸς οἰδή λόγος. ἀλλὰ δὴ ὀσπερ ἀρχὴ τούτῳ ὑποκείσθω. οὔτε γὰρ ἰατρὸς σκοπεῖ ἐι δεῖ ὑγιαίνειν ἢ μή, ἀλλ’ εἰ περιπατεῖν ἢ μή, οὔτε ὁ γυμναστικὸς εἰ δεῖ εἰ ἔχειν ἢ μή, ἀλλ’ εἰ παλαισάιν ἢ μή, ὁμοίως δ’ οἷον ἄλλη οἰδημία περὶ τοῦ τέλους: ὀσπερ γὰρ ταῖς θεωρητικαῖς αἱ ὑποθέσεις ἀρχαί, οὔτω καὶ ταῖς ποιητικαῖς τὸ τέλος ἀρχῇ καὶ ὑπόθεσις. […]

Does virtue produce the goal, or the things toward the goal? We posit that the goal, because there is no argument (sullogismos) or reasoning (logos) of it. Rather, it must be presupposed like a starting point. For neither does the doctor investigate whether to heal or not, but whether to take walks or not; nor does the athlete [investigate] whether to be in shape or not, but whether to wrestle or not. Likewise, no other [science] reasons about the goal; for just like in the theoretical [sciences] the hypotheses are starting points, equally also in the productive [sciences] the goal is a starting point and a hypothesis. […]

Thus, the starting point [archē] of thought [noûs] is the goal, and the starting point of action is the end-point of thought. Therefore, if either virtue or reason is the cause of all correctness, and reason is not the cause, then virtue will be the cause of the correctness of the goal, but not of the correctness of the things toward the goal. (EE II.11 1227b22–30 … 32–36; cf. ENX.8 1178a16–19)

The latter remarks fit nicely with the intellectualist view (see Fig. 1.2): reason stands in the middle between the goal and the action, so that the goal is reasoning’s starting point, and once the agent is done with the reasoning she can start acting. However, such intermediary role also implies that reason is not in control of its own starting point, i.e. the goal. Just like axioms in geometry, practical reason must accept the character-defined goal as a hypothesis, and take it from there, because reasoning cannot be used to produce its own starting point in the practical realm either.

(B) Deliberation is not about the goals

This is why, as Aristotle goes on to argue in the next few lines, decision (prohairesis) is concerned only with things toward the goal, but is never concerned with the goal itself.18 Decision, being a

18 “ἐστι μέντοι ἡ προαίρεσις οὐ τούτου, ἀλλὰ τῶν τούτου ἐνεκα. [However, decision is not of it [i.e. the goal], but of the things that are for its sake.]” (1227b38–39; cf. ENIII.2 1111b26–29)
product of deliberation,¹⁹ can have only the things toward the goal as its object, because, as Aristotle repeatedly claims, deliberation itself can determine only the things toward the end, and not the end itself:

βουλευόμεθα δ’ οὐ περὶ τῶν τελῶν ἄλλα περὶ τῶν πρὸς τὰ τέλη, οὐ γὰρ ἰατρὸς βουλεύεται εἰ ἱγνάει, οὐ γὰρ ἑττώρ εἰ πείσει, οὐ γὰρ πολιτικὸς εἰ εἰνομᾶν ποιήσει, οὐδὲ τῶν λοιπῶν οὐδεὶς περὶ τοῦ τέλους· ἄλλα θέμενοι τὸ τέλος τὸ πώς καὶ διὰ τίνων ἔσται σκοποῦν·

We deliberate not about the goals but about the things toward the goals. For neither does the doctor deliberate whether to cure, nor does the orator whether to persuade, nor does the politician whether to make good laws, nor do people deliberate about the goal in the other cases; but having posited the goal they investigate the ‘how’ and the ‘through what’ it will come to be. (EN III.3 1112b11–16)²⁰

So there is no reasoning about the starting points of practical reasoning, i.e., the goals; and consequently deliberation, i.e. practical reasoning, is about the things toward the ends rather than the ends themselves. But if not through reason, then how are goals determined? This will be explored in more detail in the next section (§1.3), but it will be useful to mention some passages in which Aristotle makes his answer more explicit.

**(C) Habituation determines the goals**

First, as we have already seen, Aristotle mentions virtue of character rather than reason or thought as responsible for establishing the starting points. We should see character, rather than virtue, as the crucial factor in goal-setting: a person’s character, virtuous or not, sets her goals; the virtuous character makes those goals right.²¹ So how is it that an agent’s character determines her goals?

A crucial part of the answer is Aristotle’s claim that virtue of character is about pleasures and pains.²² As seen above, an animal’s natural intuitions, given to it from birth, quite rigidly determine their practical perception, i.e. the tendencies for some things to appear pleasant and

---

¹⁹ Cf. EN III.2 1112a13–17, III.3 1113a9–12.

²⁰ The claim reappears a few lines below: “ἡ δὲ βουλὴ περὶ τῶν αὐτῶν πρακτῶν, αἱ δὲ πράξεις ἄλλων ἐνέκα. οὐ γὰρ ἐν ἐκα βουλευτόν τὸ τέλος ἄλλα τὰ πρὸς τὰ τέλη [Deliberation is about the actions that [a human being] can do, and the actions are for the sake of something else. For what is object of deliberation is not the goal, but the things toward the goal.]” (EN III.3 1112b32–34) Other explicit instances of it can be found in EE II.10 1226b10–12 and 1227a6–13.

²¹ As is mentioned in some passages, virtue (of character) preserves the goal and the starting point, whereas vice corrupts it (e.g. EN VI.5 1140b12–20, VII.8 1151a15–19).

²² The claim is repeatedly made and expanded upon in EN II.3 and EE II.2, II.4.
some things painful via their perception and phantasia. Contrastingly, human innate dispositions are much less definitive. This is not to say that our practical perception is a blank slate: we have initial natural dispositions quite similar to those of other animals. But a crucial difference between animals and humans is that our practical perception can be largely re-shaped and re-moulded through habituation, so that we end up being pleased or pained by the things that we get used to being pleased or pained by, and turn out to enjoy the activities that we have constantly repeated and become familiar to us.

The difference between animals and humans regarding habituation is important, so let me dwell on it for a moment. Aristotle does not consider animals to be devoid of all habituation capacities. His view is rather that “the other animals live mostly by nature, and some only few times by habits [τὰ μὲν οὖν ἄλλα τῶν ζῴων μᾶλλον μὲν τῇ φύσει ζῇ, μικρά δὲ ἐνια καὶ τῶις ἔθεσιν]” (Pol. VII.13 1332b3–4). Humans are born with natural dispositions that lead them mostly—and excessively—toward pleasure and away from pain (EE II.5 1222a36–38), but also have natural dispositions toward virtuous traits like justice, bravery or temperance (EN VI.13 1144b1–17). We share these initial dispositions with animals, to the point that Aristotle claims the souls of children “do not differ at all, so to speak”, from the souls of animals (HA VII.1 588a32–b2), and is prepared to attribute phronēsis and other virtue-related terms to animals that display exceptional intellectual capacities.

Animals who admit of a certain level of habituation include sheep, who can be habituated to act as bellwethers for their flock, leading them back to the barn at the shepherd’s call (HA VI.19 573b25–27); mother seals, who gradually habituate their young to spend time in the water (VI.12 567a6); and mother deer, who lead their young toward safe places to habituate them with respect to where to take refuge (VIII.5 611a19–21). Several things are worth noting: first, animals can be habituated by humans or by themselves, as the examples indicate; second, animals who partake in learning and teaching are “those which partake in hearing, not only those who perceive the differences between sounds, but also between their significations”.

23 I follow Balme’s reordering of the books in HA (Balme & Gotthelf 2002, 1–ff.).
25 “Some [animals] participate in a certain learning and teaching, some of them among themselves, some others from humans. They are those which partake in hearing, not only those who perceive the differences between sounds, but also between their significations [Ἔνια δὲ κοινωνεῖ τινὸς ἅμα καὶ μαθήσεως καὶ διδασκαλίας, τὰ μὲν παρ’ ἄλλοις, τὰ δὲ καὶ παρὰ τῶν ἀνθρώπων, ἐνεπεμφ. ὑποκείμενα μετέχει, μὴ μόνον ὡς τῶν φύσων, ἀλλ’ ὡς καὶ τῶν σημείων διαφοράς.] (HA VIII.1 608a17–21). — I take it that by “a certain learning and teaching” Aristotle is referring here to the learning and teaching of habits.
Their ability to grasp significations makes these animals able to be tamed by humans, and Aristotle thinks “tamed animals are better by nature than wild ones, and it is better for all of them to be ruled by a human; for that way they find their preservation.” But Aristotle seems to think that animals by and large need not develop habits; this makes sense given that their natural dispositions provide them with sufficient cognitive adaptation to their environments. Human innate dispositions, on the other hand, are quite deficient: we are born less hard-working and more pleasure-loving than is good for us, and our natural predispositions toward virtue are actually dangerous without the proper cognitive development.

Hence the need for habituation processes that build our character and at the same time shape our appearances of the good, therefore establishing the broad goals of our lives: what type of life appears pleasant and worth pursuing, what kind of life appears painful and avoidable. A proper habituation will lead to the correct starting points (i.e. an appearance of the good that coincides with what truly is good); any other habituation will lead to incorrect starting points (EN III.4). Thus the character we have developed turns out to establish the principles of our action: our cognitive reactions to the practical world.

Moreover, habituation not only establishes our goals by moulding our practical perception, but also determines what contemporary psychologists call our cognitive dispositions, i.e. the dispositions that determine in which situations, and to what extent, we are willing to solve problems through the use of reasoning and attend to someone else’s arguments, or rather go with what intuitively feels right and tend toward an uncritical acceptance of our own opinions. This is crucial for an assessment of the merits of intellectualism, since it reveals that for Aristotle the use of practical reason is not a characteristic of all humans, but rather only of the humans that have been properly habituated in the first place.

\[ \text{ὁ δὲ λόγος καὶ ἡ διδαχὴ μὴ ποτ’ οὐκ ἐν ἄπαισιν ἴσχειν, ἀλλὰ δὲι προδιειργάσθαι τοῖς ἔθεσι τὴν τοῦ ἀκραστοῦ ψυχῆν πρὸς τὸ καλῶς χαίρειν καὶ μισεῖν, ὡσπερ γῆν τὴν θρέψουσαν τὸ σπέρμα.} \]

26 ἄρ γὰρ ἡμερὰ τῶν ἄγριων βελτίω τὴν φύσιν, τούτῳ δὲ πάσι βέλτιον ἁρχεῖται ὑπ’ ἄνθρωπον τυχάνει γάρ σωτηρίας ὅτι οὐκ ἐν ἅπασιν ἱσχύει. (Pol. I.5, 1254b10–13, cf. Prob. X.45 896a2–3) This remark suggests that Aristotle did not spend much time in a circus or a factory farm, and that his animal-research practices probably were very humane. For further discussion of animal rule by humans, and the differences between this and other types of rule, see Miller Jr. (2013).

27 On the former, see EE II.5 (1222a36–38); on the latter, see (EN VI.13 1144b1–17). Cf. also Pol. VII.13 (1332a39–b3). Not all humans have the same innate natural dispositions: they vary in accordance with their bodily structures and climactic environments, so that different peoples have different natural dispositions toward e.g. courage and intelligence (see Pol. VII.7 1327b18–38, and Leunissen 2012 for discussion). I further discuss the shortcomings of innate human dispositions in 1.3.1 below.

28 For more on cognitive dispositions, and their relevance for a theory of agentive control, see Chapter 5 below.
οὐ γὰρ ἐν ἰκοίναις λόγου ἀποτρέποντος οὐδ᾿ αὖ συνεῖη ὁ κατὰ πάθος ζῶν· τὸν δ᾿ οὕτως ἔχοντα πῶς οἶδ’ ἐν τε μεταπείσεις; ᾧδως τ’ οὖ δοκεῖ λόγῳ ὑπείκειν τὸ πάθος ἄλλα βίοι. δεῖ δὴ τὸ ἄθανον προϊστάρχων ποιεῖν ὕπερ στερέων τῆς ἀρετῆς, στεργόν τὸ καλὸν καὶ δυσχεραῖον τὸ αἰσχρόν.

Reason and teaching certainly do not have strength in everyone; rather, it is necessary for the hearer’s soul to have been prepared with habits to be pleased and suffer finely, just like soil that is to nourish the seed. For he who lives in accordance with passion would not listen to reason, and would turn away from it—nor would he even understand it. And how is a man so disposed supposed to change his persuasion? He does not seem to yield at all to reason, but to force. So it is necessary that a character that is somewhat familiar to virtue, that loves the fine and is disgusted by the ugly, pre-exist. (ENX.9 1179b23–31; cf. Topics [Top.] I.11 105a2–7)

Character, therefore, determines not only which things one loves and hates, but also whether one listens to the voice of reason or simply lives in accordance with one’s passions. Aristotle refers to this also to argue that not everyone is poised to be a good student of politics. Immature people, who live by their passions, lack the experience that gives the proper starting points of practical reason, and that are necessary to turn reasoning into action. Arguments do not affect their actions, because they were not habituated to the use of practical reason.

To sum up, there are at least three strong lines of textual evidence suggesting that reason cannot set goals: the starting points of practical reasoning cannot themselves be obtained by reasoning; practical reasoning (i.e. deliberation) is not about the goals, but only about the things toward the goals; and our phantasiai of the good (and therefore our goals) are set by the habituation processes that constitute our character. Moreover, practical reason is able to affect our actions only to the extent that we have been habituated to listen to its voice.

RePLYING TO SOME INTELLIGENTIAL OBJECTIONS

Anyone with strong intellectualist intuitions would be uncomfortable with the way this is going. One crucial motivation for feeling uneasy is that, if reason does not play a role in goal-setting, then the whole business of establishing goals turns out to be entirely a-rational and unintelligent. Aristotle would have left the all-important starting points of action to the blind mechanism of habituation, and this simply does not sound right. The other motivation is that anti-Intellectualism seems to turn the Ethics themselves into superfluous exercises. For aren’t those texts largely dedicated to reasoning

29 ENI.3 1095a2–10, I.4 1095b4–9.
about goals? Why would Aristotle have dedicated such theoretical effort to a task that his own theory considers impossible? Moreover, he himself says repeatedly that ethical reflection is for the sake of acting, and not for the sake of knowing. This does not make sense.

About the first concern, I agree with Moss in that it seems to stem from a dualistic view according to which there are intelligent and reason-dependent processes on the one hand, and mechanical and blind processes on the other. Chapter 3 below is dedicated to demonstrating that such dualism does not hold for our current understanding of human cognition; I will argue in this Chapter (§1.3) that it does not hold as an interpretation of Aristotle’s agency theory either. The second concern, however, is much more powerful. I will argue later on (§1.4) that reason plays a crucial role in goal-setting. But right now I want to argue that the traditional explanations of how reason does this do not work, mainly because they imply bending the texts by denying Aristotle’s division of labour between character and practical reasoning.

There is a large number of controversies surrounding this topic, and settling them would require a discussion whose length exceeds what this Chapter can contain. So I will be content if in what follows I can provide a plausible case for the view that reason cannot set goals. And this is a much more manageable endeavour, for which sophisticated interpretative tricks are not required, because although Intellectualism has dominated exegetical production for some decades now, it does have a very hard case to make, and must bear a quite heavy burden of proof. After all, in its strictest form,

the intellectualist case requires that an agent can form a conception of goodness through reasoning without appeal to values already desired by the agent. The idea is that the conception so formed produces desires in the agent rather than reflecting desires she already has. For instance, reasoning to the conclusion that acting courageously is constitutive of goodness is supposed to instil in an agent who has no prior desire to act courageously a desire for courageous action for its own sake” (Grönroos 2015).

Indeed, if reason’s role in goal determination is somehow subsidiary to some character-defined motivation, then anti-Intellectualism would have the upper hand. So an argument to the effect that reason can produce cognition and motivation ex nihilo (i.e. without reference to prior habituated

30 In Moss’ (2012, 158) terms, this view stems from a conflation between the rational and the cognitive. But given that perception and phantasias are discerning (and therefore cognitive) faculties, the conflation is unwarranted.

31 Due to concerns about length, I mention below only some of the most representative intellectualist replies in the literature. For recent and much more thorough discussion, see Moss (2012, Chapter 6–7; 2014) and Grönroos (2015).
cognition and desire) would be the strongest possible case for Intellectualism. But the radical Intellectualist claim that an agent can produce appearances of the good, or desires for the good, or conceptions of the good, *without* appealing to prior non-intellectual desires or appearances or experiences, is very much at odds with the textual evidence just summarized: the division of labour between character and reason; the claim that ends are starting-points and there is no reasoning or deliberation about them; the view that character and experience, rather than *logos* and *logismos*, originate our goals and appearances of the good; and the claim that reason is ineffective in persuading those who have not been properly habituated. (I will refer to these collectively as the *division-of-labour passages*). There is, in fact, a large body of evidence that the intellectualist must contest or re-interpret in order to hold her view. A view that I will call *radical Intellectualism*.

Not so strong, but still quite powerful, would be a demonstration that reason significantly determines the goals that character only very vaguely selects. Although this form of Intellectualism (call it *moderate Intellectualism*) is also at odds with the division-of-labour passages, it is argued to have textual evidence in its favour. And despite the notorious difficulty of denying the face-value reading of so many passages, many scholars have attempted to do so, often motivated by the intellectualist concerns mentioned above.

In what follows I will examine three Intellectualist arguments, with the view in mind that if there is a way to interpret the evidence provided by the Intellectualist that harmonizes with the division-of-labour passages, this is enough to favour their straightforward reading. The arguments I will assess are (A) that through its goal-specification function reason *constitutes* those goals; (B) that *wish*, being a specifically rational kind of desire, allows us to rationally set goals for ourselves; and (C) that the grasp of practical starting-points can be performed by the intellectual capacity called *noûs*.

*(D) Specification as constitution*

Of the three kinds of intellectualism described above (§1.2.1), only specification intellectualism seems to motivate Intellectualism. It is a part of the instrumentalist view to concede that reason cannot determine the goals of action (which is why it is called ‘Humean’). And although coordination intellectualists may argue that reason defines the overall shape of the agent’s goals by the consistency constraints it puts on them, they would also agree that the shape given by reason is nothing but a refinement of the prior shape that character had already provided the agent’s goals with. Consider someone whose character makes her a lover of money: reason, in its coordinating role, might stop her from engaging in certain activities (say, taking a small earning now instead of a larger amount later), but the overall shape of the agent’s goals still corresponds
to the goals that her character produces: a rationally consistent lover of money is still a lover of money. Reason can coordinate only the material that character gives it to coordinate; its coordination function cannot provide new goals of its own. Thus, instrumental and coordination intellectualists would seem ready to endorse Burnyeat’s view that “a mature morality must in large part continue to be what it originally was, a matter of responses deriving from sources other than reflective reason” (1980, 73).

Specification intellectualists tend to be different. They can argue that deliberation’s task is more than just determining extrinsic things toward the goals (like means or consistency relations among them)—it also determines the intrinsic aspects, i.e. what exactly the goal is in particular situations. Specificationists tend to distinguish extrinsic means (e.g. killing a cow in order to obtain a winter covering from its skin) and “constitutive means” (e.g. return the money you owe in order to act justly, or stand firm in battle in order to act bravely). They know that this implies rejecting Aristotle’s explicit division of labour between character and reason. And they are prepared to do just that, by considering it an unfortunate phrasing that may lead the reader astray, or an outright mistake on Aristotle’s part.32

There are several strategies to justify this interpretative line. The most common one is arguing that the goals that character sets up are too abstract, and therefore need rational calculation in order to make them practicable. Wiggins (1975), for instance, contrasts “technical” (i.e. means-end) and “non-technical” (i.e. specificatory) kinds of deliberation, and in discussing the latter he holds that

[i]n the non-technical case I shall characteristically have an extremely vague description of something I want—a good life, a satisfying profession, an interesting holiday, an amusing evening—and the problem is not to see what will be causally efficacious in bringing this

32 McDowell, for instance, holds that Aristotle risks obscuring his own view by making the division-of-labour claim (1998a, 30). Irwin claims that it is “at least misleading” (1975, 576), because “practical intellect is not concerned with means as opposed to ends. Insofar as it is concerned with constituent ‘means,’ it is also concerned with ends” (1975, 571). This ultimately leads him to correct Aristotle’s interpretation of his own view: “[I]f a wise man’s grasp of the end is the result of his deliberation […], then Aristotle is wrong to claim that there is no reasoning about ethical first principles” (578). According to Cooper, he was “not always careful enough” to specify the right relationship between desire and intellect, “opting sometimes for this easy, but misleading, contrast” between moral and intellectual virtue (1975, 64). The trend remains strong among scholars like e.g. Frede, who calls the division of labour an “artifice” (2013, 23).
about, but to see what really qualifies as an adequate and practically realizable specification of what would satisfy this want. (1975, 38)33

Because of the extreme vagueness of the character-set goals (they amount to such things as ‘doing well’, ‘the noble’, or ‘happiness’), practical reasoning must step in to actually “determine the content of the correct conception of living well” (McDowell 1998a, 27).

But one may wonder: are character-determined goals so extremely vague as to require deliberation in order to acquire some minimally specific content? I contend this cannot be right.34 First, Aristotle compares the practical starting points with those of the theoretical sciences. These starting points must be sufficiently defined to make the subsequent deductive processes possible: geometric axioms, definitions, and postulates must have sufficient content to allow for all of Euclid’s theorems to be produced. Similarly, Aristotle holds that habituation into a certain way of acting gives agents the that sufficient to produce subsequent arguments about practical matters, along with their ensuing actions. And, again, these practical starting points cannot be inferred from deliberation, since there’s no reasoning of starting points.

Further, when Aristotle provides examples of practical starting points, they sometimes tend to be much less vague than simply ‘happiness’, ‘doing well’, or ‘the fine and the noble’. Consider one of the passages cited above, in which the doctor does not deliberate about whether to heal, and the gymnast does not deliberate about whether to be in shape—healing and being in shape taken in this case as deliberation’s starting points (EE II.11 1227b22–30).35 Earlier on (1227a13–15) Aristotle had also mentioned wealth and pleasure as potential deliberation starting points. If things like wealth and pleasure count as character-determined goals, then the view that reason is required to constitute these goals loses its strength. Of course, there is still a lot of specification to do to take the broad goal of pleasure and turn it into a particular action, but the goal’s main content is already fairly constituted by the time reason steps in. This is good news, in the end, since we do not have to engage in interpretive sleight-of-hand, or outright rejection of evidence, in order to deal with the division-of-labour passages. For they strongly suggest that reason, even in its specificationist role, does not constitute the goals’ content, but rather “merely makes explicit what is already

---

33 Cf. Irwin: “if we have a desire as vague as the desire for ‘the good for man’ or ‘the final good’ or ‘happiness,’ our deliberation cannot be purely technical, finding ways to achieve the clearly identified object of desire. It is reasoning about the end” (1975, 572). And McDowell: “The choices that display character are choices for the sake of doing well (eu prattein […]), but “to say that the end is doing well singles out nothing the agent can here and now undertake, because the question is what doing well here and now would be.” (1998a, 26)

34 In this I follow Price’s (2011) criticism of the specificationist view.

35 See also a parallel EN passage (III.3 1112b12–16) according to which the doctor’s goal of healing, the orator’s goal of persuading, and the politician’s goal of making good laws are not up for deliberation.
contained in the appearance, and thus in no way affects what goal we pursue (although it makes all
the difference to how we pursue it)” (Moss 2012, 155).36

(E) Wish, the rational desire
Another defence of Intellectualism gravitates around the concept of wish (boulēsis). The
Intellectualist can argue that reason sets goals by producing wishes, which are specifically rational
desires.37 Now, any interpretation that assumes that wish is produced by deliberation, or any kind
of reasoning process, faces the dead-end of the division-of-labour passages: goals are not
determined by logismos. But wish offers a promising alternative route for radical Intellectualists
willing to defend the claim that wish is “a basic, and unreasoned, desire of the reason-possessing
part of the soul” (Grönroos 2015, 62). If this originary rational desire can contribute to shaping an
agent’s goals, then Intellectualism may be compatible with at least the division-of-labour texts that

36 The Intellectualist can reply by pointing to evidence that the anti-Intellectualist would have trouble
accommodating. I would like to mention how some of the crucial texts may be reconciled with anti-Intellectualism.

The claim that “It seems it is proper of the practically wise person to deliberate finely about the things that are
good and convenient for him, not piecemeal […] but with respect to the good life in general” (EN VI.5 1140a25–28)
is introduced dialectical passage (as a report of someone else’s view, which is evidenced by the “it seems”), and so
should not be taken at face value. Even if it is, one may argue that “the things that are good and convenient” refer to
the things toward the goal, and “with respect to the goal” should be read as ‘with respect to the goal’.

The claim that “the good deliberator without qualification is he who is skilful in aiming at [stochastikos] what is
best for the human being in accordance with reasoning” (EN VI.7 1141b12–14) seems to connect being able to
deliberate properly well with a proper determination of the goal. But the passage does not really support
Intellectualism, since ‘being good at aiming at’ something is different from determining that something. In fact, the
thing must be already sufficiently determined before someone can successfully aim at it.

The most challenging passage for the anti-Intellectualist is perhaps the following:

εἰ δὴ τῶν φρονίμων τὸ εὖ βεβουλεῦσθαι, ἢ εὐβουλία ἢ γὰρ ὡς κατὰ τὸ συμφέρον πρὸς τὸ τέλος, οὗ ἡ φρόνησις ἀληθῆς ὑπολήψις ἐστὶν.

So if having deliberated well is proper to the practically wise person, good deliberation [euboulia]
seems to be correctness about what is convenient with respect to the goal, of which phronēsis is a
correct supposition [hupolēpsis] (EN VI.9 1142b31–33)

The first thing to remark is that the passage is ambiguous, because “of which [οὗ]” can be equally read as referring
back to “the goal” (as the Intellectualist would want) or to “what is convenient with respect to the goal”. So an anti-
Intellectualist reading is readily available. Furthermore, even if we accept the passage to state that phronēsis is the
correct supposition of the goal, the concept ‘supposition’ [hupolēpsis] must still be interpreted. Moss (2012, 180–182)
has argued on the basis of another passage (EE II.10 1226b20–30) Aristotle may take hupolēpsis to mean “the
recognition that one is working towards a given end”. If this is so, then yes, phronēsis entails having a correct hupolēpsis
of the goal, but nothing here implies that phronēsis is a reasoning about the goal, or otherwise constitutes the goal.

37 For accounts that mention the nature of wish as supportive of Intellectualism, see e.g. Irwin (1975, 569–571);
recent versions of wish-Intellectualism are defended by Grönroos (2015) and Hämäläinen (2015, 91–92).
hold that there is no reasoning about the starting points. Grönroos has recently defended this radical version of Intellectualism, according to which reason establishes desires that are independent from any prior non-rational motivation or reasoning process. (The variety of Intellectualism derived from specification-intellectualism is of a moderate kind, since the specification processes were dependent on prior processes of habituation.)

Is this kind of wish present in Aristotle’s texts? The nature of wish is far from clear. For one, it seems to be located sometimes in the rational part of the soul, and sometimes in the non-rational part. However, that is not as crucial for the wish-Intellectualist as something for which we have much better evidence: Aristotle holds that wishes are for the goals as opposed to the things toward the goals, and that they are for the good as opposed to the apparent good.

Grönroos takes these two elements and constructs a very enlightening account of wish. After

38 The most popular conception of wish among Intellectualists seems to be that it is a product of deliberation (see e.g. Irwin (1975, 570–571) Wiggins (1975, 41)). But this view hits the obstacle that there is no deliberation about goals. Grönroos’ recent defence of radical Intellectualism interestingly avoids that obstacle, which is why I focus on it here.

39 For the former, see Topics [Top.] (IV.5 126a13). For the latter, see Pol. (VII.13 1334b17–25). A controversial DA passage (III.9 432b5-6) is used by Grönroos (2015) as evidence for locating wish in the rational part, and by Moss (2012, 161–162) as evidence against it; it all depends on whether one stresses the beginning or the ending of the passage. — There is a passage that seems to support the non-rational-part side of the debate, and has not been mentioned in connection to the discussion: in EE II.4, Aristotle supports his claim that virtue of character is about pleasures and pains by attributing desire to only the non-rational part of the soul:

ἐπειδὴ δύο μέρη τῆς ψυχῆς, καὶ αἱ ἄρεται κατὰ παύτα διήρηται, καὶ αἱ μὲν τοῦ λόγου ἔχοντος διανοητικαί, ἀν ἐργον ἀληθεία, ἢ περὶ τοῦ πῶς ἔχει ἢ περὶ γενέσεως, αἱ δὲ τοῦ ἀλόγου, ἔχοντος δ’ ὀρέξιν (οὐ γὰρ ὅτιον μέρος ἔχει τῆς ψυχῆς ὀρέξιν, εἰ μεριστῇ καὶ τῷ ἄλογῳ τὸ ἐφόσον καὶ σπουδάζει εἶναι τῷ διάκειται καὶ Ἰδεῖται ἡδονᾶς τινι καὶ λύπας.

Since there are two parts of the soul, and the virtues are divided according to them, and some belong to the one that has reason and are intellectual (whose function is the truth concerning how things are or how they came to be), whereas the others belong to the non-rational part, which has desire (for not any part of the soul has desire, if it is divided in parts), it is necessary for character to be bad or excellent due to pursuing and fleeing from certain pleasures and pains. (1221b27–34)

Thus whereas truth is the only function of the rational part’s virtues, Aristotle locates all desire-related virtues in the non-rational part. (Notice that this does not entail that all cognitive functions correspond to the rational part; but rather that the rational part has only cognitive functions. The non-rational part could also have cognitive functions—which makes sense if animals are to find their way through the world only on the basis of non-rational capacities.) He further adds that not all parts of the soul have desire; and having mentioned only the rational/non-rational division, this suggests he is implicitly denying the rational part has desires. This passage has not been discussed in relation to the location of wish in the geography of the soul. At any rate, this question does not settle any issues surrounding wish-based Intellectualism, since wish is still considered to be a rational desire, its location notwithstanding. Thus, although I tend to agree with the view that all desire belongs to the non-rational part, for the sake of argument I concede Grönroos the point that wish belongs to the rational part, and contend that it still does not provide the support Intellectualism needs.

40 For the claim that wish is a desire for the end, see EN III.2 1111b26-9; III.4 1113a15; III.5 1113b3. For the claim that wish (as opposed to appetite and spiritedness) is a desire for the good as opposed to the apparent good, see EN III.4 1113a15–16; EE II.10 1227a28-31; Top. VI.8 146b5-6; Rhet. I.10 1369a2–4.
sketching his interpretation, I argue that despite its strengths the emerging picture of wish is not sufficient to ground the Intellectualist claim.

Wish is for goals rather than for things toward the goals; but what goals is wish for? Aristotle says that we wish for many things, e.g. health or victory, and even things that are impossible, like immortality, or not achievable by our own action, like success for some actor or athlete. However, despite the fact that so many things can be the object of wish, evidence strongly suggests that one particular thing is its *proper* object. In fact, given that wish is a desire for the good, if we take ‘the good’ to mean ‘happiness’ (as it often does), the specific object of wish turns out to be human happiness.

To support this view, Grönroos highlights the distinction between the natural and unnatural objects of wish, which Aristotle draws by analogy to medicine. EE II.10 presents the claim that the natural goal of medicine is health, even though a perversion [*strophe*] can be performed on it, in which case it pursues a goal contrary to nature. “And similarly—Aristotle holds—, also wish is by nature for the good, but contrary to nature it is also for the bad, and the good is wished by nature, but contrary to nature, due to perversion [*diastrophē*] the bad is also wished [ὁμοίως δὲ καὶ ἡ βούλησις φύσει μὲν τοῦ ἀγαθοῦ ἔστι, παρὰ φύσιν δὲ καὶ τοῦ κακοῦ, καὶ βούλεται φύσει μὲν τὸ ἀγαθόν, παρὰ φύσιν δὲ καὶ διὰ στροφῆν καὶ τὸ κακόν]” (1227a28-31).

This reveals that there is a natural teleology of wish: it has one specific object that is naturally proper to it, that it pursues by nature, even though some perversion can make it pursue some other object. The Nicomachean discussion of wish (*EN* III.4) provides further clarification: in it Aristotle tackles the puzzle whether the object of wish is the good or the *apparent* good. This is a puzzle because either option seems to have undesirable implications: if we admit that the object of wish is the good, then people who choose things that are not good (e.g. intemperates who choose to have excessive pleasures) are not really wishing at all; and if we admit that wish is for the apparent good, then it follows that there is no natural object of wish, and the object of wish is just whatever happens to appear good to anyone. To solve this *aporia*, Aristotle argues:

> ἆρα φατέον ἀπλῶς μὲν καὶ κατ’ ἀλήθειαν βουλητὸν εἶναι τάγαθον, ἕκάστω δὲ τὸ φαινόμενον; τῷ μὲν οὖν σπουδαῖῳ τὸ κατ’ ἀλήθειαν εἶναι, τῷ δὲ φαύλῳ τὸ τυχόν [...]· ὃ σπουδαῖος γὰρ ἐκατα κρύνει ὅρθως, καὶ ἐν ἕκάστοις τάληθες αὐτῷ φαίνεται. καθ’ ἐκάστην γὰρ ἐξιν ὀνη ἐστὶ καλὰ καὶ ἥδεα, καὶ διαφέρει πλείστον ἰσως ὃ σπουδαῖος τῷ τάληθες ἐν ἕκάστοις ὁρᾶν, ὡςπερ κανῶν καὶ μέτρων αὐτῶν ᾗν.

---

Should it be said, then, that the good is the true object of wish without qualification, but for each one it is what is apparent \(\text{to phainomenon}\)\? So that for the excellent person it is what is according to the truth, and for the base person it is any random object \([…]\); because the excellent person discerns each thing correctly, and what is true in each situation appears \(\text{phainetai}\) to him. For the fine and pleasant things are proper to each character \(\text{hexis}\),\footnote{In the ethical treatises, the word \textit{hexis}, usually translatable as ‘disposition’ or ‘state’, can receive the more specific translation ‘character’. In his discussion of the definition of character virtue (\textit{EN} II.5), Aristotle wonders whether its proper genus should be ‘affection’ \(\textit{pathos}\), ‘capacity’ \(\textit{dunamis}\) or ‘disposition’ \(\textit{hexis}\). He discards ‘affection’ for many reasons, including that we are not praised or blamed for having an affection, but rather for being affected in a certain way. He discards ‘capacity’ because we have capacities prior to their actualization, but virtues are acquired through their exercise. So \textit{hexis} remains. Thus, in action-theoretic contexts, \textit{hexis} is translatable as ‘character’, since it is determined as the genus of virtue, and virtue is a particular kind of character.} and perhaps the excellent person distinguishes himself mostly in that he sees what is true in each case, himself being a standard and measure. (\textit{EN} III.4 1113a23–33)

Aristotle’s solution to the puzzle is therefore stating that since the virtuous person is the one in whom the good and the apparent good coincide, although for everyone the object of wish is what appears good to them, nevertheless the true and natural object of wish is the good. This much is clear, but the nature of wish remains puzzling. After all, how can wish be at the same time \textit{only} for the good, and \textit{also} for anything that happens to appear good to anyone? Grönroos argues that in order to make sense of the two faces of wish, we must distinguish between the two elements that make it up as an intentional state: on the one hand, as a source of motivation, wish naturally aims at the human good; and on the other, as working on the basis of a representation of the good, wish directs motivation toward whatever appears good to the agent.

Grönroos argues convincingly that as a source of motivation wish is dependent neither on prior non-rational motivations nor on prior reasoning processes. Rather, it is an originary source of motivation: a basic, natural desire for the human good. This is why the existence of wish is good news for Intellectualism: reason sets the goals that are set by a basic, rational desire for the good.

However, in many cases (in fact, in most of them according to Aristotle) wish does not aim at the \textit{real} human good, but rather at other objects. Grönroos accounts for this by arguing that, although as a \textit{source of motivation} wish is always for the good, as a \textit{representation of the good} it may go astray (because of some perversion). Thus if a vicious person may represent pleasure as the human good, wish will motivate her toward the object that she represents as good. The key point, Grönroos suggests, is that even in cases in which the agent is pursuing the wrong object, wish is still motivating the agent toward the \textit{real} good. The error is not motivational, but representational.
This leads to the question: where does this representation of the good come from? The Intellectualist must argue (as indeed Grönroos tries to do) that the non-virtuous agent’s problem is a matter of reasoning: they just have not reached the correct conclusions about the true nature of the human good.\footnote{Grönroos explains that wish is a difficult desire to achieve (more difficult than, say, appetite for food and drink) because its object, i.e. the human good, is not perceptually available, and “requires a representation based on reason’s cognitive resources” (2015, 82). This strikes me as problematic. Reason may be required for a fully-articulated conception of the good such as we may expect from practically wise agents; but Aristotle claims wish is at work even before agents have developed their rational capacities.} But mistakes about the apparent good are not a matter of bad reasoning: as the division-of-labour passages attest, they are a matter of bad habituation. Aristotle’s explicit view, solidly attested by said passages, is that each agent’s representation (i.e. *phantasia*) of the good comes from the processes of habituation that determine her practical perception of the world’s objects as attractive or repellent, worthy of pursuit or of avoidance. The passage just quoted is explicit about this: “the fine and pleasant things are proper to each character”.\footnote{Hämäläinen objects that this passage does not establish a causal connection between character and appearance of the good. Instead, “Aristotle’s claim may only establish a correlation. He may mean that the better one’s character is, the more reliably one’s sensations indicate the goodness of a potential end, although only fully virtuous people discern good ends entirely rightly.” (2015, 108) This may be true for the passage in question. But Aristotle explicitly states a causal connection in other division-of-labour passages (cf. §1.2.2), and in passages from *EN* III.5 to be discussed below (§1.3).}

Thus, although Grönroos provides powerful evidence that “the reason-possessing part of the soul has basic desires of its own, which cannot be reduced to, or derived from, the desires of the non-rational part” (2015, 85), this remains insufficient to ground the Intellectualist view that reason can set an agent’s goals, because wish, although motivationally independent, is not cognitively independent from character and the non-rational part of the soul. Wish is therefore not a source of rational goal-setting: it is a source of rational motivation for whatever the agent was habituated into representing as good. It can accomplish its natural function (motivating the pursuit
of the human good) only in tandem with a proper habituation, because habituation still sets the appearance of the good. All of which implies that character determines the goal even of wish.\textsuperscript{45}

In reply, the wish-Intellectualist could adduce evidence that even people who were badly habituated, and now have a false representation of the good, still wish for the real human good and feel inner conflicts due to their apparent good not matching what really is good.\textsuperscript{46} This may be so, but it still falls short of demonstrating wish-Intellectualism, which would require reason to provide remorseful vicious people with a character-independent representation of the goal. Otherwise, wish provides no rational, character-independent action guidance. If reason was able to provide such guidance, we would expect to find in Aristotle's texts mentions of cases of reflective conversion, i.e. cases in which vicious people correct their representation of the good through philosophy or some other rational activity.\textsuperscript{47} But such cases are not forthcoming. What is forthcoming is Aristotle's view that only those who have been properly habituated can profit from reasoning about practical matters, and that legislators should not try to reason with those badly habituated, as one of the division-of-labour passages (\textit{ENX}.9 1179b23–31) suggests, since most people listen to their passions, like fear and pleasure, but not to reason. What should they do instead?

\begin{verse}

\textit{ἐκ νέου δ’ ἀγωγῆς όρθῆς τυχεῖν πρὸς ἀρετὴν χαλεπὸν μὴ ὑπὸ τοιούτου τραφέντα νόμοις· τὸ γάρ σωφρόνως καὶ καρτερικῶς ὕδω τοῖς πολλοῖς, ἄλλως τε καὶ νέος. διό νόμοις δεῖ τετάχθαι τὴν τροφὴν καὶ τὰ ἐπιτηδεύματα· οὐκ ἔσται γὰρ λυπηρὰ συνήθη γενόμενα. οὐχ ἢκανόν δ’ ἵσως νέους ὄντας τροφῆς καὶ ἐπιμελείας τυχεῖν ὀρθῆς, ἀλλ’ ἐπειδὴ καὶ ἀνδρωθέντας δεῖ ἐπιτηδεύειν αὐτὰ καὶ ἐβίβαζειν, καὶ περὶ ταῦτα δεοίμεθα· οὐκ ἔσται γὰρ λυπηρὰ συνήθη γενόμενα. διὸ στρέφεσθαι τοῦ καλοῦ ὧν ἂν νόμοι ῥήμα καὶ ἐπιτηδεύουσι καὶ ἦμαις τῷ καλῷ. διόπερ οἴονται τινες τοὺς νομοθετοῦντας δεῖν μὲν παρακαλεῖν ἐπὶ τὴν ἀρετὴν καὶ προτρέπεσθαι τοῦ καλοῦ χάριν, ὥς ἐπακουσομένων τῶν ἐπιεικῶς τοὺς ἐθήσει προηγμένων.
\end{verse}

\textsuperscript{45} Grönroos, following the Intellectualist tradition, tends to over-intellectualize the human grasp of the good. He claims that the good is too complex an object to be an object of perception. But the \textit{ENIII}.4 passage just quoted shows the contrary: in it, Aristotle says that the excellent person “distinguishes himself mostly in that he \textit{sees} [\textit{hor} \textit{ān}] what is true in each case”. Further, the passage is crowded by uses of the verb \textit{phainesthai} (and nothing suggests they should not be interpreted technically as references to the faculty of \textit{phantasia})—but does not use a reason-related verb even once.

\textsuperscript{46} The evidence for this is Aristotle’s claim that “probably even in base people there is some natural good stronger than their own [baseness], which desires for their proper good \[τοσοῦ δὲ καὶ ἐν τοῖς φαύλοις ἐστὶ τι φυσικὸν ἄγαθὸν κρεῖττον ἣ καθ’ αὐτά, ο ἐφίεται τοῦ ἑαυτοῦ ἄγαθοι]” (\textit{ENX}.2 1173a4-5), along with the passage that states vicious people have no friendly feelings for themselves, but rather a great remorse due to the strong internal conflict between their appetite and their wish (\textit{ENIX}.4 1166b7-22). Cf. Grönroos (2015, 79–80).

\textsuperscript{47} Cases like that are narrated by Grönroos: “take a person who thinks that excessive bodily pleasure is the ultimate good, and that the sybaritic life will make her happy. If eventually she achieves this kind of life, but finds out that she is not content, and is perhaps even bored and disgusted with the kind of life she leads, she may try to revise her view of what the good life comes to and of what kind of life makes her happy.” (2015, 81)
ἀπειθοῦσι δὲ καὶ ἀφυεστέροις οὖσι κολάσεις τε καὶ τιμωρίας ἐπιτιθέναι, τούς δ’ ἀνάτονοι ὅλως ἐξορίζειν· τὸν μὲν γὰρ ἐπιεικῆ πρὸς τὸ καλὸν ζῶντα τῷ λόγῳ πειθαρχήσειν, τὸν δὲ φαῦλον ἡδονῆς ὀρεγόμενον λύπῃ κολάζεσθαι ὡσπερ ὑποζύγιον. διὸ καὶ φασί δεῖν τοιαύτας γάνθαι τὰς λύπας αἱ μάλιστ’ ἐναντιοῦνται ταῖς ἀγαπωμέναις ἡδοναῖς.

But it is difficult for people to come across the right education towards virtue from childhood if they have not been brought up by adequate laws, because the life of moderation and perseverance is not pleasant for the many, particularly for the young. Hence it is necessary to prescribe upbringing and customs by law; for things that have become habitual will not be painful.

And it is perhaps insufficient to hit upon the right upbringing and care while being young; rather, since it is necessary to practice and be habituated into these things when they are mature. And we need laws concerning that, and in general concerning all of life. For the many obey necessity more than reason, and punishment more than the fine.

This is why some think that the legislators must, on the one hand, exhort people towards virtue and persuade them for the sake of the fine, since those who have been brought up with decent habits will listen; and on the other hand, impose punishments and penalties on those who are disobedient and bad-natured, and banish those who are completely incurable. For the decent person, living with a view to the fine, will listen to reason, whereas the base person, desiring pleasure, is to be punished just like a beast of burden. Hence they also say it is necessary to bring about the pains that are most opposed to the pleasures he loves. (EN X.9 1179b31–1180a14)

Rational persuasion is thus reserved for those who have the required precondition: a proper habituation. Base people are not expected to change their ways via reason, but via punishments similar to those used to tame animals. The lack of cases of rational conversion therefore seems to support anti-Intellectualism.

In sum, although there is evidence in favour of the claim that wish is a basic, rational, character-independent source of motivation, wish cannot work without the representational content provided by character. So no solid foundation for Intellectualism is to be found here.

(F) Practical nous

There is another suggestive alternative for those who admit there is no deliberation about ends as such, but still want to say that grasp of the ends is rational. They can consider the grasp of the end to be some sort of intellectual intuition. Aristotle helpfully has a name for this intellectual,
albeit non-demonstrable, kind of cognition: *noûs*. Cooper’s *Reason and human good in Aristotle* (1975) provides a compelling analysis in this direction, so it is worth assessing it now.

Cooper admits there is no deliberation about the ultimate end, and concedes that the grasp of the good comes about through habituation, rather than intellectual instruction. This leads him to the view that the virtuous person’s knowledge of the ultimate goal “must be a kind of intuitive knowledge, not based on reasons of any kind” (1975, 62). What worries Cooper about this is that a *phronimos*, if questioned about his conception of the good, may be unable to provide an articulable justification for it. There should be a way for the virtuous person to argue in favour of his view of the ultimate end; but if not through practical reasoning, then what kind of reasoning could he employ?

Drawing a parallel between the theoretical and practical realms, Cooper argues that *dialectic* may be adequate for the job. In the theoretical sciences, dialectic is the argumentative method whereby Aristotle thinks we can gain a grasp of a science’s non-demonstrable first principles. Dialectic argument is more informal than the scientific deduction: its raw materials are common observation, opinions, or linguistic usage. Since there is no demonstration of a science’s first principles, they are said to be known by *noûs*, i.e. intellectual intuition.

Arguably, the first book of the *EN* is a dialectical argument that seeks to establish the first principles concerning the human good. If so, the virtuous person can justify his views about the good, if not through practical reasoning (i.e. deliberation), then at least on dialectical grounds. Thus, Cooper concludes, the principles concerning the good are grasped by *noûs*; and they can be established by dialectical argumentation (1975, 60–65).

I find Cooper’s interpretation very compelling, and mostly in line with what I have been proposing here. I would only distance myself from the first conclusion, i.e. that we grasp the good by *intellectual* intuition (or *noûs*). One may think that this intuition has to be intellectual insofar as it has been enriched with the dialectical analysis performed in *EN* I. And that is probably correct, but remember that by the time someone is exposed to Aristotle’s arguments about the good she is already supposed to have had a correct grasp of the *what*. The *EN* I dialectic is probably directed at providing the *why*, but only on those grounds already provided by habituation. Because the habituation-based grasp of the *what* is prior to any dialectic, the dialectic justification cannot provide evidence to consider the *what* rational. The only way to

---

48 This conception of dialectic is based by Cooper on Owen’s (1961) classical article on the topic.

49 Cooper mentions *Posterior Analytics* [*APo*] I.2 (71b16–17); I.3 (72b18–20); II.19 (100b5–15) as evidence. One may add Aristotle’s claim that there is *noûs*, but not *logos*, of both the first and last things (*EN* VI.11 1143a35–b5). For the discussion surrounding that passage, see Moss (2012, 189–190).
argue that this habituation-based cognition of the good is intellectual would be to say that habituation itself is an intellectual process of some kind. Cooper—and many others—think that this is the case, but I find no evidence for such an interpretation of habituation. I will argue below that habituation is a completely non-rational process (§1.3.1); but for now, I argue that Cooper himself should agree with the view that the originary grasp of the good is not intellectual, given his own theoretical commitments.

In his discussion, Cooper distinguishes between the question “whether for any agent […] the commitment to an ultimate end is based on reasons”, and the question “whether or not considerations can be advanced which tend to show that a given end […] is the best end to treat as ultimate” (60). Although he gives a positive answer to the second question (dialectical argumentation can produce considerations in favour of treating a certain end as ultimate), he thinks Aristotle must answer “No” to the first question, given that people’s grasp of the ultimate end originates not in reasoning, but in habituation. Thus, he concludes:

It is prudent and reasonable, I think, of Aristotle not to insist that ordinary moral agents know, or need to know, how to defend their own view of what kind of life is best: their intuitive conviction is a perfectly adequate foundation for them. But those dialectical and critically alert persons who are moved to deepen their understanding of this foundation can do so by examining, in a dialectical spirit, the deliverances of moral intuition. (1975, 71)

I think this is entirely correct: human agents (even those with a correct grasp of the good) need not have a sophisticated theoretical articulation of the good. To do this, one must engage dialectically with the subject—but that is an entirely different matter from becoming habituated. The what of habituation is not the why of dialectic. So even if the why provides an intellectual grasp of the good, there is no reason to think that the what is also intellectual.

The role of reason in action production: is that it?

The intellectualist may be upset about the minimal role this account gives to reason in the constitution of goals. And rightly so, because some aspects of Aristotle’s doctrine remain unaccounted for. Consider Aristotle’s claim that desire should obey reason.50 This seems to go against the role that I have given to reason up until now, as nothing but an intermediary

50 In the passage that explains the divisions of the soul, Aristotle claims the desiderative part of the soul partakes in reason “insofar as it listens to and obeys reason ἣν κατήκοον ἐστίν αὐτοῦ καὶ πειθαρχικών” (EN I.13 1102b31; cf. EE II.1 1219b29-1220a2). For more on this hierarchical relationship, see §1.4 below.
between desire and action. In fact, if reason cannot determine goals, but is rather constrained by
them, and they are shaped by non-rational cognition, reason seems to be obeying desire, and not
the other way around.

Moreover, the worry about what the Ethics are for remains unanswered. After all, Aristotle’s ethical project is from the beginning a rational study of the goals of life, and Aristotle is quite explicit that we study ethics “not so that we may know what virtue is, but so that we may
become good” (EN II.2 1103b27–8). But how can we become good by studying ethics, if reasoning cannot contribute to shaping our goals?

These are both important concerns. But the division-of-labour passages are too strong to be ignored or bent into one’s favourite form of Intellectualism. I think none of the alternatives so far reviewed are satisfactory: we cannot rest content with an anti-Intellectualist account of action production for the aforementioned reasons, but the traditional Intellectualist ways are foreclosed by the division of labour. So it is necessary to find another way. In what follows I will argue that, although direct Intellectualism (according to which reasoning can directly affect one’s goals) is indefensible, there is however an indirect form of Intellectualism that is compatible with the division of labour between character and reason, and able to account for the crucial role that reasoning plays in becoming and being good.

But an exposition of indirect Intellectualism requires a few intermediate steps. To start heading in that direction, notice that the main question I have been asking (what makes it the case that humans have a greater control over their actions than the other animals?) remains unanswered. Particularly, we still do not know how we may be in control of our goals, the very principles of our action. Because of this, it could be argued that we still have not accounted for human agentive control, since having no control over the starting point of action in a way entails that we have no control over the process as a whole. So how can we say that we are in control of our actions, if we are not in control of our appearance of the good? And how can we be in control of our appearance of the good, if we are not in control of the mechanisms that determine it? These are crucial questions.

Luckily, Aristotle has an explicit answer.

---

51 Both concerns are taken as reasons to support Intellectualism: since Aristotle clearly thinks that intellectual development plays a crucial role in becoming good, and anti-Intellectualism is unable to account for that role, anti-Intellectualism must be rejected. Hämäläinen (2015, 108) has recently criticized Moss’ revival of anti-Intellectualism along these lines.
1.2.3. Aristotle’s solution to the problem of goal control

In his Nicomachean discussion of action (ENIII.5), Aristotle addresses an opponent who objects that while virtuous people are voluntarily virtuous, no one is voluntarily vicious. The beginning of this intricate passage makes it clear that the stakes of this debate are very high: if it cannot be demonstrated that vice is voluntary, “we should not say that man is a principle or that he is the generator of his actions” (1113b17–19). In other words, either Aristotle successfully refutes his opponent or his whole theory of action falls apart. The issue is very serious indeed, for unless it can be shown that virtuous people are voluntarily virtuous and vicious people are voluntarily vicious, the very claim that we are in control of our actions is falsified. This is due to the tight causal connection there is between actions and character: a certain character begets actions of a certain kind (1114a11–18) because the principles of our actions, i.e. the appearances of the good that establish their goals, are determined by character (1113a30–31). So we are the principle of our actions only if our we are the principle of our character too.

So Aristotle presents the opponent’s case:

εἰ δὲ τις λέγοι ὅτι πάντες ἐφίενται τοῦ φαινομένου ἀγαθοῦ, τῆς δὲ φαντασίας οὐ κάριοι, ἀλλ’ ὁποῖος ποθ’ ἐκαστός ἐστι, τοιούτο καὶ τὸ τέλος φαίνεται αὐτῷ. [...] οἴθεις αὐτῷ αἴτιος τοῦ κακοποιεῖν, ἀλλὰ δ’ ἔγνωκαν τοῦ τέλους ταύτα πράττει, διὰ τούτων οἴόμενος αὐτῷ τὸ ἁριστόν ἐσεθαι, ἢ δὲ τοῦ τέλους ἐφεσίς οὐκ αὐθαίρετος, ἀλλὰ φῦναι δὲ ἄφθερ ὠφεν ἔχοντα, ἢ κρατεῖ καλῶς καὶ τὸ κατ’ ἀλήθειαν ἀγαθὸν αἰρήσεται, καὶ ἔστιν εὐφύς ὃ τούτῳ καλῶς πέφυκεν.

Now, if someone said that everyone pursues the apparent good, and we are not in control of phantasia but rather the end will appear to each one in accordance with the sort of person that he is, [...] no one would be responsible for his own bad actions, but rather each one would do what he does due to ignorance of the goal, believing because of it that what he does is the best for himself. And the aim for the goal would not be self-chosen; instead, one would need to have been born with a natural sense of sight that allowed one to discern

52 The text provides no explicit identification, but there seems to be a consensus among scholars that this opponent is a Socratic who takes Socrates’ no-one-does-wrong-willingly doctrine and derives from it that vice is not voluntary, although virtue and happiness are. (It has even been suggested that it is Socrates himself (Boeri 2008, 10).) The opponent’s claim, however, is expressed in a poetic style (οἴθεις ἐκῶν ποιητοὶ οὐδ’ ἂκων μακάρος (ENIII.5 1113b14–15)) that resonates with the work of Epicharmus, an author who predates Socrates (see Gauthier & Jolif (1970, 213); Irwin (1999, 208); Broadie & Rowe (2002, 321)).

53 Aristotle’s own reply appears in the middle of this phrase, interrupting the opponent’s argument. I bracket it for a moment in order to present the opponent’s argument in full.
finely and choose what is truly good. And he who has this sense in a fine condition by birth
is a good-natured man. \(EN\text{III}.5\ 1114a31\text{–}b5\)

Taking ‘\(φαντασία\)’ and ‘\(φαινόμενον ἀγαθόν\)’ in the technical sense (i.e. as referring, respectively, to
the faculty of \(phantasia\), and to the appearance of the good that this faculty generates),
the objection can be reconstructed as follows:

(I) Our \(phantasia\) of the good is determined by the sort of person that we happen to be.

(II) It is not up to us to be the sort of person we happen to be.

(III) Therefore, our \(phantasia\) of the good is not up to us. [I, II]

(IV) So if someone has the wrong \(phantasia\), she will act badly, and not voluntarily. [III]

(V) Therefore, no one is voluntarily vicious. [IV]\(^{54}\)

The opponent’s argument has an interesting positive flipside: because we are the kind of person
that we (naturally) turn out to be, our \(phantasia\) of the good is not up to us, but up to \textit{nature}. So
if we end up doing the wrong actions, it is because we have the wrong \(phantasia\), and this in turn
is due to our having a bad nature. On the other hand, those who have the correct \(phantasia\) are
in a way naturally blessed, born with a special sense of practical perception that provides them
with the right discernment tools. For the rest of us, we can at least rest content in that our errors
are not up to us, but rather caused by an unavoidable ignorance, a natural imperfection: we were
just born that way!\(^{55}\)

Aristotle could reply to the objection by rejecting either (I) or (II). If he held an
Intellectualist view, one would expect that he replied either (I\textsuperscript{*}) that our \(phantasia\) of the good is

---

\(^{54}\) The argument seems to rely on two uncontroversial implicit premises. IV can be inferred from III and the
implicit Aristotelian thesis that people act pursuing what appears good to them (e.g. \textit{EN}III.4). V can be inferred
from IV and the Aristotelian claim that a person’s character stems from the kinds of actions she repeatedly
performs (e.g. \textit{EN}II.1).

\(^{55}\) This naturalistic account of goodness appears elsewhere in Aristotle’s texts. The (rather puzzling) \textit{EE} VIII.2 is
dedicated to examining the possibility that some people may turn out to aim correctly at the \textit{goal}, \textit{and} achieve it,
out of pure luck. He argues that fortunate people (i.e. those who do well in life without ever really knowing what
they are doing) cannot actually achieve the goal of life out of fortune \(tuche\), because luck cannot produce the same
result throughout an entire life. So fortunate people (who reliably act on the basis of the right desires at the right
times and in the right ways) are successful due to their good nature rather than luck. The argument goes on to say
that the cause of someone’s good nature must be a sort of divine principle, which bypasses and surpasses reason, and
accounts for things like divination. To come back to our passage, Aristotle seems to be dealing here with the
radicalized version of the same naturalistic account of moral success. \textit{EN}III.5 is much more critical of the
naturalistic approach than \textit{EE} VIII.2, however, perhaps because of an increased awareness that if we are not in
control of our character we cannot be in control of our actions either. If our \(phantasia\) of the good depends on our
natural dispositions, how can we be in control of the actions that originate in it?
determined by reason; or (II*) that we can determine the kind of person we are through the exercise of reason. But Aristotle replies to the objection by arguing neither for I* nor for II*.

Supporting I* would, for one, imply going against the division-of-labour passages, and he does not do that in \( EN \) III.5. Nor does he reply by endorsing II*: this passage contains no hint of the claim that logos or logismos can determine a person’s character. Familiar Intellectualist views that one can reach a proper grasp of the goal via practical noûs, or that wish points us naturally in the direction of what is truly good, have no role in Aristotle’s retort. In fact, no word related to reason appears anywhere in his reply. Aristotle does reply by rejecting (II), but in the way that an anti-Intellectualist would expect:

\[
\text{εἴ δὲ τις λέγων ὅτι πάντες ἐφίενται τοῦ φαινομένου ἀγαθοῦ, τῆς δὲ φαντασίας οὐ κύριοι, ἀλλὰ ἀστεῖος πολλ' ἐκαστός ἔστιν, τοιοῦτο καὶ τὸ τέλος φαίνεται αὐτῷ: εἴ μὲν οὖν ἐκαστὸν ἑαυτῷ τῆς ἐξελίξου ἐστὶ πως αἰτίος, καὶ τῆς φαντασίας ἐσται πως αὐτὸς αἰτίος.}
\]

Now, if someone said that everyone pursues the apparent good, and we are not in control of phantasia but rather the end will appear to each one in accordance with the sort of person that he is—actually, if each one is in a sense the cause \([aιτίοι]\) of his own character \([hexis]\), he will also be in a sense the cause of his phantasia. (1114a31–b3)

I translate the word aitios here as ‘cause’ to highlight the causal aspect of the claim, but it could also be translated as ‘responsible’ or ‘accountable’, and both things would also drive the point home: you are accountable for your appearance of the good because, and insofar as, you are the cause of your own character \([hexis]\).56

The obvious question here is: how are we the cause of our character? Answering this question requires an account of how the process of habituation and moral development works.

---

56 Compare this link between phantasia and character \([hexis]\) with the \( EN \) III.4 claim mentioned earlier: “the excellent person discerns each thing correctly, and what is true in each situation appears \([phasisetai]\) to him. For the fine and pleasant things are proper to each character \([hexis]\)” (1113a29–31). See Curren (1989) for an examination of the relationship between moral and causal responsibility.
§1.3. Habit and the structure of human agency

If we control our actions only insofar as we are causes of our phantasia of the good, and we are causes of the latter only insofar as we are causes of our character, then in what sense are we causes of our character? I will try to answer this question by (1.3.1) providing an account of how the processes of habituation, moral, and cognitive development work. Once this account is in place, (1.3.2) the overall structure of human agency will become clearer: character, practical cognition, desire, reason, and action mutually shape each other in a process of circular causality. This makes it possible for us to be causes of our character by acting in a way that develops the character we want.

1.3.1. How habituation works

The notion of habituation is deceptively simple: according to Aristotle, we become habituated into being pleased and pained by the right things simply by performing virtuous actions (EN II. 1). But habituation then seems to be a process of rote repetition. If it is, then it does not seem able to explain the generation of the complex cognitive capacities and dispositions that make an integral part of adult virtuous life (like good deliberation and phronēsis). Because the process of repetition-based habituation seems too mechanical to account for intellectual development, some scholars have included a conceptual or intellectual aspect in their description of habituation, according to which the process is not only about repeating certain tasks and developing certain dispositions, but is also a dialectical training grounded on exhortations and advice, and its upshot is a set of beliefs about the noble, the shameful, and the like, and a capacity to apply these concepts to particular situations and actions. From this perspective, the fruits of proper habituation (the that Aristotle considers a prerequisite for political education) are correct beliefs about the fine and the noble, and a conception of the human good,
perhaps implicit in these beliefs, that is ready to be thematized in the study of ethics. In fact, intellectual development seems to some interpreters to be so central that moral education has even been said to consist merely in acquiring the intellectual capacities required for *phronēsis* and the other intellectual virtues, and integrating them with the already adequate non-rational dispositions we have from birth.

I will argue that it is not necessary to intellectualize habituation in order to account for cognitive development. Aristotle’s discussions of habit and habituation do not explicitly mention intellectual or conceptual elements, but only repetitions of actions. But if we understand it correctly, a non-intellectual, dispositional account of habituation is sufficient to explain how we develop the proper intellectual capacities that ultimately lead to the theoretical virtues if all goes well.

In this section I will provide an account of habituation by looking first at the relationship between habit and nature, and then at how exactly it is that mere repetition generates cognitive changes in our *phantasia* of the good. I will argue that (A) habit is an artificial inner causal principle that extends our natural inner principles; (B) and it works because through repetition we become able to endure affections and activities we are initially unable to endure. Given this account, we can move on to explain how (C) moral and (D) intellectual development grow out of this repetition-based process: with the proper habituation, we become able to engage in, and ultimately enjoy, the effortful activities that ultimately lead to the virtues.

---

57 Cooper thinks Aristotle, though not “careful enough to say so”, does not conceive of habituation as “the purely mechanical thing it may at first glance seem”; what is not mechanical about it is that it “must involve also […] the training of the mind”, so that the trainee “comes gradually to understand what he is doing and why he is doing it”, i.e. “to see the point of the moral policies which he is being trained to follow, and does not follow them blindly” (1975, 8). Burnyeat, for his part, holds that the difference between only having the *that* and also having the *because* consists in “a contrast between knowing or believing that something is so and understanding why it is so”, and that the necessary starting points of practical reasoning are “correct ideas about what actions are noble and just” (1980, 71–72). In a similar vein, McDowell claims that “We travesty Aristotle’s picture of habituation into virtue of character if we suppose the products of habituation are motivational propensities that are independent of conceptual thought, like a trained animal’s behavioural dispositions. On the contrary, the topic of book 2 is surely initiation into a conceptual space […] organized by the concepts of the noble and the disgraceful […]. Possessing ‘the that’, those who have undergone this initiation are already beyond uncomprehending habit […]. They have a conceptual attainment that, just as such, primes them for the reflection that would be required for the transition to ‘the because’.” (1998a, 39–40). Likewise, Frede holds that “habituation concerning the virtues of character […] does not only consist in acquiring the disposition to be affected correctly in every particular situation, but also the disposition to choose the right action. And both involve a good deal of thought.” (2013, 23)

58 There is important textual evidence for this view: in his brief discussion about natural vs. full virtue (*EN* VI.13), Aristotle seems to state that children already come equipped from birth with appropriate character dispositions, and thus all that they are missing for full virtue is acquiring *noës*. On these grounds, Lennox (1999) argues that moral development consists in intellectual development plus the integration of the newly developed rational capacities with the natural virtues of character.
What is habituation?

(A) Habit as an extension of nature

This is one of Aristotle’s descriptions of habituation:

ἐστὶ τὸ ἰθὸς, ὀσπερ καὶ τὸ ὄνομα σημαίνει ὅτι ἀπὸ ἐθους ἔχει τὴν ἐπίδοσιν, ἐθίζεται δὲ τὸ ὑπ’ ἀγωγῆς μὴ ἐμφύτου τῷ πολλάκις κινεῖσθαι πώς, οὕτως ἢδη τὸ ἐνεργητικὸν.

Character [ἔθος], as its name indicates, is that which grows out of habit [ἔθος], and something is habituated when, after having been repeatedly moved in a certain way by non-innate training, it then enacts that pattern of activity itself. (EE II.2 1220a39-b3)

Habits have a causal force: through repetition an agent becomes able to perform a non-natural pattern of activity without external enforcement. At first, the principle that accounts for the agent’s activity is external to the agent, but through habituation the principle is internalized. Producing the activity thus quite literally becomes ‘second nature’: an artificial inner principle of movement.

Habit’s role in memory confirms the parallel between habit and nature: in On Memory and Recollection [Mem.], Aristotle says we recollect whenever we move from one starting idea to the idea we wish to remember. We do this by creating a ‘memory path’ that leads from the starting idea to the target idea. But whenever an “old path” has not been created, “one’s movement tends to the most habitual; for habit now assumes the role of nature. This is why we remember quickly the things we think about often; because just like one thing follows another in nature, it is also like that in the actualization [of memories]; and repetition produces nature.” With this striking sentence—“repetition produces nature”—Aristotle seems to mean that repetition generates an inner causal principle analogous to nature, i.e. habit.59

As mentioned before, the inner animal principles of motion are cognition and desire; and habit also has cognitive and motivational aspects. Habituation informs these inner principles by making something appear as pleasant or painful. “The things that are customary and acquired by habit are also among those that are pleasant—Aristotle says in the Rhetoric—; for many among the things that are not naturally pleasant produce pleasure when people have

59 ἄν οὖν μὴ διὰ παλαιοῦ κανηβῆ, ἐπὶ τὸ συνηθέστερον κινεῖται· ὀσπερ γὰρ φύσις ἢδη τὸ ἰθὸς, διά τι πολλάκις ἐννοούμεν, ταχὸς ἄναμμασακάμεθα· ὀσπερ γὰρ φύσει τὸ μετὰ τὸ ἔστιν, οὕτω καὶ ἐκερχείς· τὸ δὲ πολλάκις φύσιν ποιεῖ. (Mem. 2 452a27–30) On the habit-nature relation, see also EN (VII.10 1152a30–33), Magna Moralia [MM] (1203b31–04a2).

Notice the similarities between this process of habitual recollection and the ‘automatic-associative’ mental processes discussed in contemporary debate (see §2.2 below). Similar worries arise with respect to both: is this a purely unintelligent, rote-based process? If so, how could we be in control of habitual actions? (See Chapters 2–3 below.)
become habituated to them.” Since pleasure and pain are properties of perceptual objects, and thus the crucial links between cognition and desire (§1.1.2), habits are cognitive and motivational principles. Aristotle’s HA discussion of human sexual development exemplifies both causal and cognitive aspects of habit:

So through repetitions of (both physical and mental) activity we generate habits that shape not only the subsequent motions of our body, but also our cognitive and motivational dispositions, i.e. our tendencies to perceive things as appealing or unappealing, and therefore to generate desires towards them. This way, habit is an extension of nature, in the sense that it extends the activities of our inner principles of practical phantasia and motivation beyond what nature’s innate dispositions do. Habit extends nature because by repeating the same kinds of activities we create new patterns of practical cognition and motivation, whose characteristics were not innately established (e.g. if we repeatedly engage in sexual intercourse, we subsequently desire sexual intercourse more than was innately determined).

60 ἔστιν δὲ καὶ τὸ σύνηθες καὶ τὸ ἐθιστὸν ἐν τοῖς ἡδέοις· πολλὰ γὰρ καὶ τῶν φύσει μὴ ἡδέων, ὅταν συνεθισθῶσιν, ἡδέως ποιοῦσιν· (Rhet. I.10 1369b16–18). For further discussion of the nature-habit analogy, see also EN II.3 (1104b8–13), VII.10 (1152a30–33), and Corcilius (2013, 141–142).
(B) Repetition, endurance, and pleasure

But how is it that repetition of actions can shape our cognitive and motivational dispositions toward feeling pleasure and pain? This is the core problem of habituation: repetition seems to be a purely mechanical process, so it is not clear what cognitive value it may have. How can we move from mechanical iteration to cognition and motivation? Luckily, Aristotle explicitly answers this question.

The first step towards the answer is understanding the link between what is natural and what is pleasant. We have found this link before, in Aristotle’s claim that “what is in accordance with nature is pleasant” (HA VIII.1 589a2–9; cf. §1.1.1). The Rhetoric further develops the idea:

Let us suppose that pleasure is a movement and a complete perceived settling of the soul toward its original nature, and pain its opposite. […] Thus what for the most part leads toward what is according to nature is necessarily pleasant, and most of all whenever the beings that are according to nature recover their own nature[.] (Rhet. I.11 1369b33–1370a5)

In this passage Aristotle assumes the traditional view (associated with the Hippocratic school and depicted in several Platonic dialogues, like the Philebus and the Gorgias) according to which pleasure is the perception of a process whereby a living being recovers its natural state of harmony. (This model explains e.g. why animals desire food when hungry, and find eating pleasant in that condition, but do not desire food when satisfied.) From this perspective, what is pleasant is tightly connected with what is natural. But what is natural, or leads toward a natural state, does not exhaust the list of pleasant things. Aristotle adds two more elements:

καὶ τὰ ἔθη (καὶ γὰρ τὸ εἰδισμένον ὀσπερ πεφωκός ἤδη γένεται· ὁμοιῶν γὰρ τὸ τὸ ἔθος τῇ φύσει· ἐγγὺς γὰρ καὶ τὸ πολλάκις τῷ ἁρι, ἐστιν δὲ ἡ μὲν φύσις τοῦ ἁρι, τὸ δὲ ἔθος τοῦ πολλάκις), καὶ τὸ μὴ βίαιον (παρὰ φύσιν γὰρ ἤ βία, διὸ τὸ ἀναγκαῖον λυπηρόν [...])

As is well known, in his two treatises on the subject (EN VII = EE VI 11–14 and EN X.1–5) Aristotle rejects this explanation of pleasure—which he explicitly accepts here only as a supposition—and provides two (not entirely consistent) theories that link pleasure with the exercise of an activity of the soul rather than with the recovery of a lost state. That said, the link between pleasure and nature is arguably still present in Aristotle’s own account, since activities of the soul are actualizations of our natural psychological capacities (cf. e.g. EN VII.14 1154b15–20).
Also habits [are necessarily pleasant] (for what has been habituated already occurs just like what is natural, because habit is something similar to nature; in fact, what occurs repeatedly is akin to what occurs always, and while nature is among the things that occur always habit is among those that occur repeatedly). And what is not forced [is necessarily pleasant] (for what is forceful is against nature, which is why things that are necessary are painful [...] (1370a5–10)

This passage confirms the nature–habit analogy just explained, and builds an overall picture of the things that are pleasant, which are split into three groups. What is natural (because it restores a natural state), what is habitual (because habit works in a way analogous to nature), and what is not forced (because what is forced is opposed to what is natural). The third group actually includes the other two, so it is granted to say that anything that is not against nature, i.e. forced, is pleasant.62

This is the first element required to answer our question (how can repetition alter our cognitive dispositions?). The second element is that, although those things that are not against nature are pleasant, they are pleasant only to a degree. Eating and drinking, for instance, are pleasant until you are full, and after a certain point those activities can become quite unpleasant. Similarly, we enjoy physical and cognitive activities (like walking or running, learning, or contemplating), but only until we get tired.63

This is where repetition comes in: repeating the same activity increases our capacity to endure its performance, thereby making us able to enjoy it for longer. This is made explicitly clear in Aristotle’s reply to one of the Problems concerning food, which is worth quoting in length:

Διὰ τι τὰ αὐτὰ συνεθιζομένοις τε ἢδεα φαίνεται καὶ λίαν συνεχῶς προσφέρομένοις οὐχ ἢδεα; τὸ δὲ ἐθὸς ἐστὶ τὸ πολλάκις καὶ συνεχῶς τι ποιεῖν. ἢ ὅτι τὸ μὲν ἐθὸς ἐξεν δεκτικὴν τυχον ἐν ἡμῶν ποιεῖ, οὐ πλήρωσιν, τὸ δὲ συνεχῶς προσφέρομεθαί τι πληροὶ τὴν ἐπιθυμίαν, καὶ καθάπερ αἴτιον; ἐστι γάρ τι καὶ ἡ ἐπιθυμία, αἱ μὲν οὖν ἡξες γυμνάζομεν αὐξάνονται καὶ ἐπιδιδόσαν· τὰ δὲ ἀγγεῖα συνεχῶς ποιεῖν μείζω γίνεται. διόπερ τὸ μὲν ἐθὸς ὃν

1. Habit and reason in Aristotle’s theory of agency — 41

62 An animal moves by force (βίᾳ) whenever it moves or stays still against its own inner impulses, because of the influence of an external source of motion (EN III.1; EE II.7). (We can imagine a hurricane lifting a cow and dragging her through the sky—whenever that happens, it happens by force—, or a cage impeding a bird’s flight—its stillness occurs by force, against the animal’s inner sources of motion—). Aristotle’s point in this Rhet. passage seems to be that an animal (particularly, a human) can enjoy anything that is not imposed upon it by external principles, because it can naturally perform or receive it.

63 πῶς οὖν οὐδὲς συνεχῶς ἢδεται; ἢ κάμνει; πάντα γὰρ τὰ ἀνθρώπων ἄδυνατες συνεχῶς ἐνεργεῖν. οὐ γάνεται οὖν οὐδ’ ἡδονή ἐπεται γάρ τῇ ἐνεργείᾳ. [“How come, then, nobody is pleased continuously? Is it because they get tired? For it is impossible for anything human to be continuously active. So there cannot be [continuous] pleasure either, for pleasure follows the activity.”] (EN X.4 1175a3–6)
Why do the same things appear pleasant to those habituated [to them], but do not appear pleasant to those who take them in too continuously? Habit, however, is doing something repeatedly and continuously. Is it perhaps because habit produces in us a certain receptive disposition \[\text{hexin dektikē tēn} \]\, but not satiety, whereas continuously consuming something satiates the appetite, just like a vessel? 

For appetite also is something [like this]. Thus dispositions, when trained \[\text{gumnazomenai} \], grow and increase; but vessels do not become larger by getting filled up. Hence habit, being an exercise \[\text{on gumnasion} \], increases the receptive disposition; but that which is continuously taken in fills us up and satiates our appetite, and when our appetite is satiated we do not receive any more, and it does not grow at all by packing it full, for the reasons stated.

Further, habit is pleasant not in the sense that it always is pleasant (for even these things are painful, if they are done continuously), but rather in the sense that we receive the beginning of the activity with pleasure, and are able to do the same thing for a longer time than if we were not habituated.

In that sense, then, things produce pain even though they are pleasant, and in that same sense it occurs with the other pleasant things. For things are painful both if they come to be or are consumed continuously. The cause is that our own receptive and productive capacities are not unlimited, but limited; and when they come across something that is commensurate with them (one can indeed continuously and increasingly perceive this), some of them get satiated, and the others are unable to function. \((\text{Prob. XXI.14 928b23-929a5})\)

Given the limitations of our capacities for receiving affections and producing actions, we find pleasure in affections and activities only until these capacities are exhausted. But habit extends our natural capacities by creating receptive dispositions. By repeating an activity we train...
ourselves in its performance, building up our endurance of this activity—much like a runner builds up endurance while training for a marathon, or someone who regularly eats spicy food develops a resistance for spices. This ‘gymnastic’ aspect of habit is crucial for understanding its role in both moral and cognitive development.\(^\text{65}\)

Crucially, habit does not seem to make us enjoy things that we did not enjoy before at all—i.e. we cannot habituate ourselves into things that are forced and go against our inner principles of action. What habit does do is allow us to extend our natural pleasure by extending the amount we can receive of the pleasant thing before it becomes painful, or extending the amount of time we can perform an action before we get too tired and have to take a break. If a virtuous life is like a marathon for the soul, proper habituation is like the regular training that builds the endurance we need to reach the end of the race without hitting a wall of exhaustion.

To summarize, repeating actions shapes and extends our inner principles of cognition and motivation because the regular repetition of an activity works just like training: it allows us to endure the activity, while taking pleasure in it, for a longer period than we were initially and naturally able to. Thus, habituation is the training of the soul that extends its receptive and active capacities. Because the soul’s activities, being natural, are pleasurable in themselves, habituation also extends the pleasure we can enjoy in performing these activities. This is why proper habituation leads to the correct moral and intellectual development.

**Habitation and development**

**(C) Habituation and moral development**

Earlier I mentioned an interpretation according to which moral development was all about intellectual development, i.e. the acquisition of rational capacities like noûs and phrónesis. This view, defended by Lennox (1999), is crucially based on Aristotle’s discussion of natural virtue in the *Nicomachean Ethics*. In this passage, Aristotle holds that “we are just and naturally moderate and brave dispositions already from birth, but we nonetheless search for something else to be full goodness [τὸ κυρίως ἀγαθὸν]”.\(^\text{66}\) This makes it seem like we already have the proper non-rational dispositions in place at birth, and what we need to acquire is the proper rational capacities to

---

\(^{65}\) Similarly, in *ENII.2* (1104a27–b3) Aristotle claims that the effect of repeating the right actions is developing endurance: we become strong by enduring much hard work (πολλοὺς πόνους ὑπομένειν), and those who are strong are most capable (μάλιστα δυνάμεθα) of hard work; likewise, abstaining from pleasures makes us temperate, and temperate people are those most capable (μάλιστα δυνάμεθα) of resisting pleasures.

\(^{66}\) καὶ γὰρ δίκαιοι καὶ σοφρονικοὶ καὶ ἀνδρεῖοι καὶ τάλλα ἔχομεν ἐθνὸς ὡς γενετῆς· ἄλλ’ ὁμοιοί ζητοῦμεν ἐτερῶν τι τὸ κυρίως ἀγαθὸν. (*ENVI.13* 1144b5–7)
adequately handle them—a suspicion that is corroborated by the further assertions that such innate “natural virtues” are harmful without intellect, and that “if someone acquires intellect, he will act differently; and the disposition, though similar, will then be full virtue \([\text{kuriós areté}]\).” 67

But moral development cannot only consist in intellectual development, because our innate dispositions are not perfect; in fact, they are skewed away from the fully virtuous mean, pointing us to different kinds of excesses. Aristotle makes this clear when in EE he claims that, although too much pleasure and too much pain are both deviations with respect to the correct mean, “from the start nature does not depart from the mean similarly in every case; rather, we are less hard-working and more devoted to enjoyment [than we should be].” 68 If we had the correct dispositions towards pleasure and pain, we would not need to learn to enjoy and be pained by the right things—we would already do so. But Aristotle stresses that it is crucial to be appropriately brought up precisely for that reason: “to make us find enjoyment or pain in the right things; for this is the correct education.” 69

This is precisely what habituation is supposed to do: shape our pleasure-and-pain dispositions by training, so that we get to enjoy the right things and feel pain toward the right things, because through repetitions we extended our natural dispositions for reception and action in the right directions, so that we can perform the right actions, and enjoy them, for extended periods; but did not extend our natural dispositions in the wrong directions, so that we cannot perform the wrong actions, nor do we enjoy them, but it would rather pain us to perform them for extended periods, since we do not have extended receptive and active dispositions towards them.

Moreover, the gymnastics of habituation also explain how we learn to enjoy things that are naturally painful, and to be pained by things that are naturally pleasant—something that is necessary to become virtuous, since e.g. courage requires being pleased by enduring terrible

---

67 ἐὰν δὲ λάβῃ νοῦν, ἐν τῷ πράττειν διαφέρει· ἣ δ᾿ ἔξις ὑμιασ τότ᾿ ἡσα τις κυρίως ἀρετή. (1144b12–14) Based on this passages, Lennox (1999, 13) claims that “the move from natural to unqualified virtue seems simply to be the acquisition of practical intelligence by beings with natural virtue”. He goes on from here to argue for the Intellectualist view that the division of labour between character and reason is “utterly artificial”, and that “the virtuous character that consistently promotes the actions of a good life must be guided by practical intelligence”, implying that phronēsis must partake in establishing goals. (His view is more interestingly complex than this, in the sense that he claims that moral education’s role is to provide an “integration of practical intelligence with dispositions to feel and act, such that one’s feelings and actions tend to be appropriate expressions of the life of a rational and political animal” (25). I find this idea [i.e. that moral development consists in the integration between phronēsis and proper character dispositions] compelling. What I criticize is the idea that the proper character dispositions are with us from birth.)

68 ἢ φύσις εἶθεν οὕτως πρὸς ἀπάντημα ὑμιασ ἀφετῆκε τοῦ μέσου, ἀλλ᾿ ἦττον μὲν φιλόπονοι ἔμεν, μᾶλλον δ᾿ ἀπαλαστικοί (EE II.5 1222a36–38).

69 δεὶ ἢ καθαί πώς εἴθες ὡς νέων, ὡς ὁ Πλάτων φησίν, ὡστε χαίρειν τε καὶ λυπεῖσθαι αἷς δεί· ἢ γὰρ ὀρθὴ παιδεία αὕτη ἔστων. [“It is necessary to be somehow habituated immediately from early youth, as Plato says, so that we enjoy and are pained as we should; for that is the correct education.”] (EN II.3 1104b11–13)
threats to our own safety, and temperance requires being pleased by *abstaining* from pleasures. We can train our pleasure-and-pain dispositions in these non-innate ways only through repetition of temperate and courageous actions (*NE* II.2, 1104a33–1104b3).

If this is so, then the gymnastic conception of habituation can satisfactorily explain how we develop the right moral dispositions. But how about rational capacities? Can this account also explain how we acquire *phronēsis* and the other intellectual virtues?

**(D) Habituation and intellectual development**

Interpreters tend to intellectualize habituation in order to explain intellectual development: if the intellectual capacities of virtuous agents stem from it, habituation must be more than repetition-based non-rational training; it must also include what McDowell calls the “initiation into a conceptual space”. Of course, habituation must allow for the beginning of conceptual reasoning; but maybe habituation itself needn’t be conceptual. A non-intellectual, repetition-based process of habituation allows us to acquire intellectual virtues because it implies practicing rational activities until we expand our capacities to endure their performance, to the point that we become so good at them that their proper pleasures are sufficiently motivating to keep advancing toward excellence. The activities themselves are rational and conceptual, but the habituation and the dispositions developed through them need not be.

I will make my case for this view by tackling an associated problem: what motivates us to practice in the first place? Reasoning is difficult and requires effort that goes against our innate impulses, so it cannot be initially pleasant. How do we get motivated to practicing effortful activities in the first place, then? The *Rhetoric* chapter cited earlier might offer some clues:

\[
\text{τὰς δ’ ἐπιμελείας καὶ τὰς σπουδὰς καὶ τὰς συντονίας λυπηρὰς· ἀναγκαία γὰρ καὶ βίαια ταῦτα, ἣν μὴ ἐθισθῶσιν· οὕτω δὲ τὸ ἔθος ποιεῖ ἤδις. τὰ δ’ ἐναντία ἤδεα· διὸ αἱ ῥαθυμίαι καὶ αἱ ἀπονίαι καὶ αἱ ἀμέλειαι καὶ αἱ παιδιαὶ καὶ αἱ ἀμέλειαι καὶ αἱ ἀναπαύσεις καὶ ὁ ὕπνος τῶν ἠδῶν· οὐδὲν γὰρ πρὸς ἀνάγκην τούτων.}
\]

The careful, effortful, and intense [activities] are painful; for they are compelled and forced, unless they are habitual (but in that case, habit makes them pleasant). And their opposites are pleasant; which is why easy-goingness, laziness, carelessness, games, rest, and sleep are among pleasant things; for none of them are compelled. (*Rhet.* I.11 1370a11–15)
Recall that habituation is the process whereby we come to internally generate activities that initially require an external principle of motion (§1.3.1A). Things that initially require effort and care are forced upon us, and thus are not pleasant: they go against our natural tendency to slack and our minimal natural capacities to perform them. But by repeatedly exercising them we develop the endurance necessary to undertake them more thoroughly and enjoy them for longer periods. Through habituation, effortful attempts become unimpeded activities. And the latter are far more enjoyable!

And this is how we get to become good reasoners: by practicing, so that we become able to reason each time a little better and for a little longer. This is actually all we need, because rational activities are pleasant in themselves (being actualizations of natural capacities); so if we extend our capacity to remain active in thought, we also extend our experience of the pleasures of thinking. And as we make progress in the gym of the mind, intellectual pleasures themselves become the motivating force that leads us to exercise intellectual capacities more often. This positive cycle of practice leading to pleasure leading to practice is sufficient to explain how we become good thinkers. So non-rational habituation sufficiently explains cognitive development; there is no need to add conceptual and intellectual elements into the habituation process itself.

My view is thus in line with the dominant interpretation of motivation in moral development, according to which “the child’s sense of pleasure, which to begin with and for a long while is his only motive, should be hooked up with just and noble things so that his unreasoned evaluative responses may develop in connection with the right objects.” (Burnyeat 1980, 79–80). Jiménez has recently raised strong criticisms against “pleasure-centred accounts” like this. Her objection from priority goes like this:

[A]lthough we find natural activities innately pleasant, pleasure in a nonnatural activity is only available to those who have become familiar with that activity and have learned to love it (typically through habit). Thus, the proper pleasure that learners take in virtuous activities cannot be the basis through which learners grasp and pursue at first the goodness of those activities, precisely because only those who already engage in virtuous activities on account of their nobility or goodness are able to take proper pleasure in virtue. (2015, 138).

This account of the reasoning-habituation link has the virtue of explaining Aristotle’s views that logos is unable to motivate the many (see §1.2.2C), and that improperly habituated people cannot reap fruits from practical reasoning: reasoning is useless for all of them because they cannot follow it, since it requires more effort than they, having underdeveloped dispositions for reception and action, can exert.

Jiménez mentions Broadie (1991, 122n46) and Cooper (1999, 277) as earlier proponents of the objection from priority.
I agree with several of Jiménez’s points, but think the objection can be met by a pleasure-centred view based on the account of habituation sketched above. The first thing to notice is that Jiménez takes virtuous activities to be “nonnatural”. And she is right to do so, in a sense, because virtues do not grow in us naturally: we must be habituated into performing them. But in another sense virtuous activities are natural, insofar as they are the proper actualizations of the human soul’s natural capacities for cognition, volition, and motion. Rational activities are natural to human beings; what is “nonnatural” (in the sense of ‘requiring habituation’) is the ability to perform them excellently—that virtuous disposition must become second nature through habituation.

The objection is, thus, that rational activity requires so much effort that it soon becomes painful for the learner. So the initial motivation to engage in rational activity cannot be the activity’s pleasure, because it is too scarce, if at all present, and is quickly opaqued by discomfort, exhaustion, and pain. Those who endorse the objection from priority, however, go on to claim that “enjoying something as noble requires having a previous minimal grasp of the noble” (Jiménez 2015, 142), thus suggesting that this cognitive grasp can do what pleasure cannot: give the learner the initial inner motivation to practice rational actions. This is where I must disagree, for two independent reasons. First, the priority objection assumes that some cognition can do better than pleasure at initially motivating the learner towards noble actions. But as I argued earlier (§1.1.2), motivation and cognition are intrinsically linked: the cognition of something as good just is the appearance of something as pleasant. So grasping something as good and finding it pleasurable are inseparable. Grasping something as noble is no exception, since the noble must also appear pleasant to be motivating. If this is so, then there is no problem of priority, since grasping something as noble and finding it pleasant are simultaneous occurrences.

72 “There are some elements of the performance of each of the capacities and crafts that require prior education and habituation; so clearly this also applies with respect to virtuous actions” (Pol. VIII. 1 1337a18–21).

73 On the distinction between first and second natures, see above (§1.3.1A) and Vasiliiou (1996, 778–781).

74 Here are a couple of passages that suggest this: “The path [of moral progress] is through pleasure; and noble things must be pleasant [διὰ τοῦ ἱδέου δὲ ἡ ὁδὸς· ἀνάγκη εἶναι τὰ καλὰ ἱδέα.]” (EN VII.2 1237a6-7). — “While there are three objects of choice […]: noble, advantageous, and pleasant […], the good person is correct about all of them, and the bad person is mistaken about all of them, but mostly about pleasure; for it is common to animals, and it follows everything that occurs by choice, since also the noble and the convenient appear pleasant. [τριῶν γὰρ ἀναλοικτῶν τῶν εἰς τὰς αἱρέσεις καὶ τριῶν τῶν εἰς τὰς φυγές, καλοῦ συμφέροντος ἱδέως, καὶ τριῶν] τῶν ἔναντίων, ἀδικοῦ βλαβεροῦ λυπηροῦ, περὶ ταῦτα μὲν πάντα ὁ ἐγχώριος κατορθωτικός ἢστιν ὁ δὲ κακὸς ἁμαρτητικός, μᾶλλον δὲ περὶ τὴν ἱδίαν κοινῆ τε γὰρ ἀντίκειται τοῖς ἔναντι καὶ πάσα κατὰ τὸν ἀγαθὸν παρακολουθεῖ· καὶ γὰρ τὸ καλὸν καὶ τὸ συμφέρον ἱδίον φαίνεται.]” (EN II.3 1104b30–1105a1)
The second reason is that the grasp of the noble cannot be the inner motivation that gets the learner started in the practice of virtuous activity because no inner principle can account for these initial learning stages. If it could, habituation would be unnecessary, since its job is internalizing an initially external source of motion. Aristotle’s claim that we need to be properly habituated thus implies that the initial motivation to start learning has to come from outside. This is presumably because starting to perform noble activities requires a lot of effort and pain, so children would not do it on their own. They may be able to perform them for a little while, but would then get tired and distracted, and would ultimately be unable to reach the intensity required to extend the correct character dispositions. Therefore, children need education, i.e. careful habit guidance from competent adults who serve as external sources of motion. We need habituation because we need external support—an external principle of motion—to get our moral and cognitive development going. So Jiménez is surely right in that “Aristotle does not think that pleasure and pain function as the natural guides to virtue” (2015, 156). If pleasure was our natural guide, we would hardly end up being anything but self-indulgent and careless. But a cognitive grasp of the noble cannot play that role either. The natural guide to virtue is neither pleasure nor cognition, but other people. Only when our parents, teachers, and other caregivers have made us sufficiently able to endure, we can take the wheel and continue the process by caring for our own habits.

Now, despite the fact that pleasure is not our main initial source of motivation, it nevertheless is the main inner source of motivation throughout the learning process. The problem is not that the pleasures of rational action are not there from the beginning—the problem is that they are too quickly exhausted and replaced by pain due to exhaustion. So initially habituation

---

75 “While learning, [children] are not having fun, because learning is painful [οὐ γὰρ παίζουσι μαθάνοντες, μετὰ λάπης γὰρ ἡ μάθησις]” (Pol. VIII.5 1339a28–29). According to Diogenes Laertius, Aristotle said that the roots of education are bitter, but the fruit is sweet [Τῆς παιδείας έφη τὸς μὲν ρίζας εἶναι πικρός, τὸν δὲ καρπὸν γλυκύν].

76 The unavoidable relevance of others for the individual’s habituation reveals that the process whereby each agent acquires proper control over her actions is necessarily political (see §1.4 below for the full argument). Such initial dependence on others seems to be why Aristotle claims in EN III.5 that each one is the cause of his dispositions, and therefore of his phantasia, only “in a way [παί]” (1114b2–3). This is contra the opinion that this qualification is due to the phantasiai depending on natural dispositions beyond our control (e.g. Boeri 2008, 11–12). The point of habituation is, after all, being able to reshape those natural dispositions.

77 This should be qualified. Aristotle claims the body is the first to develop, then the desiderative part of the soul, and finally the rational part (Pol. VII.13 1334b17–25). My claim is then that young humans can take pleasure in rational activities as soon as their natural rational soul develops, since pleasures accompany their corresponding activities. This pleasure is the same that a virtuous person feels in performing rational actions. But the young agent’s pleasure is much shorter and feeble, because her capacities for action performance are equally brittle, so her action is quickly impeded. In this respect my view crucially differs from the pleasure-centred theorist Jiménez criticizes, who “says that by performing virtuous activities the learners discover (and acquire a taste for) those new pleasures of the noble” (2015, 140).
requires an external source of motion. But as habituation progresses and the learner extends her endurance, the point at which she requires the external force to keep going is pushed further, and the inner source of motivation increasingly takes the leading role. This inner source of motivation is the pleasure of rational action, which grows stronger and extends itself through time as the learner progresses. So I agree with the roles Jiménez attributes to pleasure: “Pleasure confirms that learners are indeed on the right track (it is a “sign” of their moral progress); and it encourages those able to properly perform virtuous actions by intensifying those very activities” (2015, 139). Except I do not think there is much else motivating the agent from within.78

1.3.2. The circular structure of human agency

Let us look back at the big picture for a moment. Recall the anonymous opponent who argues that we are not in control of our actions because we are not in control of our phantasia of the good (§1.2.3). In response, Aristotle claims that we are in control of our actions insofar as we cause our appearance of the good, and we are responsible for our appearance of the good as we cause our character. The claim implies that understanding how character is caused is necessary to understand control over actions. In the previous subsection (§1.3.1) I attempted to explain how habituation causes character and explains moral and intellectual development.

This puts us in a better position to understand how an agent can be the cause of her own character. A crucial part of the explanation turns out to be that human agency has a circular structure (see Fig. 1.3): repeated actions generate dispositions of character; character in turn determines practical cognition (i.e. the phantasia of the good); practical cognition activates desire toward the objects that appear good; and desire—mediated by reason’s specificatory, instrumental, and coordinating activities—shapes subsequent actions. These actions then generate dispositions of character… and the cycle starts again.

---

78 Jiménez raises the further objection (2015, 149–151) that pleasure cannot be the motivation behind performing virtuous activities because “for an action to be virtuous and virtuously performed it has to be chosen for itself” (149) and not for the pleasure that follows from it. However, the distinction between an action and its proper pleasure is arguably not as straightforward as this makes it seem. For on the theory defended in EN VII.12 the pleasure we take in performing an unimpeded activity is the activity itself. It will surely be replied that this conception of pleasure is usually taken to be less definitive than one in EN X.4, in which pleasure is no longer identical to activity, but rather completes it. Moreover, as Jiménez points out, this account has it that we would choose to perform certain activities even if no pleasure ensued (X.3 1174a4–8). But the ENX version is more ambiguous in this respect than is usually recognized. In it, Aristotle also hesitates to answer whether we choose the activities of life because of pleasure or pleasure because of these activities, “since these things [i.e. pleasure and life’s activities] seem to be yoked together and not allow for separation [οὐ καὶ δεῖχθαι ὑπ’ αὐτὸν ἔνα ἔργον]” (X.4 1175a19–20). So Jiménez’s objection that we should choose virtuous actions for their own sake and not for their corresponding pleasure relies on the dubious presupposition that pleasure and activity are separable, something that Aristotle doubts is possible. The topic deserves further treatment (which would go beyond this work’s limits), but for now I wanted to point out that there are reasons to suspect this objection may not work.
This marks a structural difference between human and animal agency, since animal action production is a linear causal process, whose starting point is the innate natural dispositions and whose end point is locomotion (see Fig. 1.1). In the case of basic agents (i.e. non-human animals and children), nature innately specifies their objects of desire through their innate phantasai of the good. So animals are in control of their motions insofar as the principles of those motions are internal to them; but they have no control over the principles of their action, because their phantasia of the good is by and large hard-wired into them by innate natural tendencies (§1.2.2C). The only way for an agent to take control of the principle of her action (i.e. the first mover, the object of desire, which sets all other movers in motion) is by making the whole system bite its tail, so that actions are able to modify their own principles. This is what a human agents can do (modify their phantasia of the good) if she can shape her habits. It thus becomes possible for us to act in order to acquire a certain character, thereby shaping the way the practical world appears to us.

Figure 1.3: The circular structure of human agency.

Aristotle’s reply to the anonymous opponent is now clear: we are in control of our actions because we are in control of our phantasia; and we are in control of our phantasia because we are in control of our habits. By transitivity, it follows that we are in control of our actions
only insofar as we are in control of our habits. All of this leaves one important question open: how is it that we have control over our habits? Should we assume that all human adults, insofar as they are such, are in control of their habits? Or is it a matter of achievement?

As I will argue now, it is an achievement. More specifically, a political achievement.

---

79 More specifically, Aristotle claims that we are “in a way responsible” [pós aitios] of our phantasia of the good and our habits (EN III.5 1114b2–3). I mentioned above that this qualification is due to the unavoidable fact that we must rely on other people for our initial habituation. So each individual agent is only one of the people responsible for her character; the others who are also responsible include her parents and teachers, and most crucially, the legislators of her city (see §1.4).
§1.4. Reason as care: A defence of indirect Intellectualism

There is no way we can set our life’s goals directly by just reasoning about them, because reason cannot directly alter our sources of motivation. This is clear from the structure of human agency. But reason could determine the phantasia of the good indirectly, by directing the process of habituation toward the development of certain character dispositions, which will in turn make certain things appear pleasant and others painful. This is indirect Intellectualism, the view that logos can determine our goals by carefully shaping our habits. In this section I offer evidence that Aristotle in fact endorsed indirect Intellectualism, and explore the implications that this has for the meaning of logos. With this, my discussion of the foundations of human agency in Aristotle will be complete.

Earlier on (§1.2.2) I mentioned some arguments in defence of intellectualism that remain standing. One of them is Aristotle’s view that reason should govern desire. This hierarchical view is widespread. When separating the rational and non-rational parts of the soul, Aristotle often makes the point that there is one aspect of the irrational part that does not have reason, but can nevertheless listen to it, like a father. The rational part is also said to play other caregiving roles. Now, if, in the economy of the soul, the role of the non-rational part is to obey, listen to, and live in accordance with reason, like a child must do with respect to their caregivers, then arguably the role of the rational part is, like a father, to care for the non-rational part. Since our control over actions depends on our control over character, then this care for habits turns out to be a necessary condition for agentive control.

Clearly, care for habits is a task that only reason can perform: it requires considering actions instrumentally as means to build a certain disposition, specifying which actions would better lead to that goal, and performing long-term coordination; so all of reason’s functions are implied. Through care for habits, reason can shape and determine our goals (so Intellectualism is correct), but not by simply reasoning about them (so traditional direct Intellectualism is not an accurate description). Reason’s control over goals is indirect in two ways. First, my reason can shape my motivations only by way of shaping my character. And second, I can take care of my character only because someone else took proper care of it first: only thanks to the rational care of others I can exercise rational care myself. This second aspect of indirect Intellectualism has important implications for the conception of logos, because, given that habituation is not entirely up to us, but rather starts from those who originally cared for us (§1.3.2D), care for habit is originally a matter of public reason, rather than private reason.

80 E.g. “As the child should live according to the direction of his tutor, so the appetitive element should live according to reason” (EN III.12 1119b3–18). See also EN I.13 1102b31, IX.4, 1166a13–14; EE II.1 1219b29-1220a2; Pol. I.5 1254a33–34, b3–6.
After (1.4.1) unpacking the argument for the claim that care for habits is necessary for action control, (1.4.2) I review the texts where he argues that care is a task that should belong to the polis rather than to individuals, and that the legislation is a public form of reason. The ultimate upshot of these arguments is that the reason of individuals depends on the public reason that structures the polis through the laws. (These latter remarks sound quite strange, I know; but I think we must face the fact that Aristotle’s thinking is quite foreign from ours in this respect, and that these political views of his affect the way he conceives of ourselves as individuals. Only by looking at the strangeness of someone else’s thought can we fully recognize—and critically assess—our own intellectual commitments.)

1.4.1. Care for habits as necessary for control

For me, running is both exercise and a metaphor. Running day after day, piling up the races, bit by bit I raise the bar, and by clearing each level I elevate myself. At least that’s why I’ve put in the effort day after day: to raise my own level. I’m no great runner, by any means. […] But that’s not the point. The point is whether or not I improved over yesterday. In long-distance running the only opponent you have to beat is yourself, the way you used to be.

—H. Murakami: *What I talk about when I talk about running*

After concluding that human agency has a distinctively circular structure, the question emerged whether for Aristotle human adults are in control of their habit and character just by being such, or whether control over habits (and therefore control over actions) was some sort of achievement. To find his answer, we must examine the remaining part of the *EN* III.5 debate with the anonymous opponent.

Immediately after introducing his opponent’s main thesis (i.e. that no one is voluntarily vicious), Aristotle begins his refutation (1113b21–30) by claiming that everyday social practices presuppose that we are in control of our actions: we all praise and blame, reward and punish each other for actions we perform—all of which would make no sense if an external principle determined our actions; so these practices attest to the fact that we are in control of our actions.

Further, our everyday legal practices show that people are sometimes held responsible *even* for things they do in ignorance (1113b30–1114a3); e.g. if someone did something in ignorance because he is drunk, he gets twice the punishment. This practice—he thinks—reveals that those who “seem to ignore things “because of lack of care [δι’ ἀμέλειαν]” are accountable for their actions, because not ignoring it was up to them; since “they were in control of caring [τοῦ γὰρ ἐπιμεληθέναι κύριον].”
Now that Aristotle has introduced the notion of care, the opponent strikes back, challenging the very idea that everybody is in control of caring, of being careful:

ἀλλ’ ἵσως τοιοῦτος ἐστιν ὅστε μὴ ἐπιμεληθῆναι.

—“But perhaps he is the sort of person not to take care!” (1114a3–4)

This remark, laconic as it is, makes a lot of sense: after all, being careful about things is a character disposition (a *hexis*), and the dispute is about whether we actually are in control of such character states. So Aristotle’s answer thus far seems to merely beg the question. But he retorts:

ἀλλὰ τοῦ τοιούτους γενέσθαι αὐτοὶ αἴτιοι ζῶντες ἀνειμένως, καὶ τοῦ ἀδίκους ἢ ἀκολάστους εἶναι, οἱ μὲν κακοκρανθοῦντες, οἳ δὲ ἐν πάσι καὶ τοῖς τοιούτοις διάγοντες· αἱ γὰρ περὶ ἐκαστὰ ἐνέργεια τοιοῦτοι ποιοῦν. τοῦτο δὲ δήλων ἐκ τῶν μελετῶντων πρὸς ἰμπυνοῦν ἀγωνίαν ἢ πρᾶξιν· διατελοῦσι γὰρ ἐνέργοισι. τὸ μὲν οὖν ἀγνοεῖν ὅτι ἐκ τοῦ ἐνεργεῖν περὶ ἐκαστα ᾧ ἔξεις γίνονται, κομιδὴ ἀναισθήτην.

—However, they are the causes of having come to be that sort of person, by living slackly [aneimenōs]. And they were the causes of being unjust, by doing bad actions, and of being intemperate, by having spent their time drinking etc. For each sort of activity produces the corresponding sort [of character disposition]. And that is made evident by those who practice [tōn meletōntōn] any competition or activity [praxin]; because they accomplish this by being active. So you have to be absolutely insensible to ignore that performing actions generates the corresponding states of character [hexeis]. (1114a3–10)

The dialectic ends here. Aristotle goes on to expand on his conclusion by stating that, therefore, people who commit unjust acts are voluntarily unjust, and no one who acts badly can say that he is involuntarily unjust. So he takes it that by this point he has already won the argument. —But how? What was his winning argument?

In order to show that being virtuous or vicious is in our control, Aristotle adduces a crucial piece of evidence: *the behaviour of athletes* (i.e. “those who care about some competition or practice”). This is rather surprising, if not disappointing. He seems to assume that merely watching someone train for a competition reveals how control over habits works. But, really, what does an athlete working out have to do with the voluntariness of vice?
On closer inspection, however, and keeping in mind the description of habituation as the soul’s gymnastics, maybe all we need to know about habit and agency is before our eyes when looking at an athlete training. In order to be able to perform at her best level, she knows she must practice, i.e. she knows she must repeat the same sort of action over and over and over, carefully, until she generates the kinds of sensorimotor dispositions she needs to be able to endure performing consistently at her peak. Because she knows the right dispositions are generated by repeating the right kinds of actions, she takes control of her habits by carefully repeating the kind of actions that will turn her into who she wants to be. So the athlete’s behaviour does reveal the secret of agentive control: we are in control of our actions insofar as we carefully shape the habits that cause them. This is a variation on the theme found in the *Prob.* passage discussed above (§1.3.1B): habituation is an exercise, a gymnastikon, through which we extend our naturally limited dispositions for receiving things and producing actions, thereby extending our control.

The fact that this dialectical passage is full of words related to care (ameleia, epimelethēnai, meletān, tōn meletōntōn) highlights its main upshot, namely that care for one’s character is a necessary condition for being in control of one’s actions. Someone who ignores this (the dependency between actions and character states) would be “absolutely insensible”, i.e. would lack all understanding of the human condition. And there could clearly be people who lack this self-understanding: Aristotle would not hesitate to deny it to natural slaves, probably to women, and perhaps even to the majority of people, who go about their lives guided by their passions and do not care at all about the fine (*EN* X.9 1179b31–1180a14). The problem for those who do not care for the shape of their habits is that they end up stuck with severely limited dispositions for sensation and action, that lead them to excessively cling to immediate pleasure and flee from even little amounts of pain. So they never develop the ability to control their actions in demanding situations: they will go for the small instant gratifications, and miss out on the hard-earned satisfactions of a fulfilled life. From this perspective, then, agency is not a property of each and every human being, given by the possession of the rational faculty. Rather, it is an achievement that is attained by continuously elevating oneself beyond one’s past levels through habit training.
1.4.2. Legislation, reason, and care

The concept of care [epimeleia] plays a crucial role in Aristotle's action-theoretic argument in response to the anonymous opponent. It reappears in a context that may seem different, but is actually continuous with this reflection: the discussion about the transition from ethics to politics. Talking about a 'transition' is a bit misleading, since Aristotle describes the study of ethics as a political endeavour from the beginning: he claims that the human good is the object of study of political science, the most controlling [kuriōtē] science that should determine how a city's resources are to be distributed among all the other sciences and capacities.\(^{81}\) It should be no surprise, then, that care for habits, being so crucial for an individual agent's capacity to control her own life, turns out to be a political issue.

Public vs. private care

The idea that the soul's rational part should be like a father to the non-rational part is not just a metaphor: parents are the first rational beings who are responsible for carefully shaping their children's habits, and parents quite literally are their reason—the external, rational principle of motion that shapes their originary habits—while their ability to rationally care for their own character develops. Later on, their tutors and educational institutions take on part of this role. Even more generally, though, caring for their citizens' character development may correspond to the city's legislators. That is, in fact, their job: “legislators make citizens good by habituating them, and that is the wish of every legislator; those who do not do this well err in their purpose, and a good constitution differs in this respect from a bad one.”\(^{82}\)

Moreover, according to Aristotle the city's legislation should be more responsible for children's habitation than the parents themselves. This is why:

\(\text{ἡ μὲν οὖν πατρικὴ πρόσταξις οὐκ ἔχει τὸ ἰσχυρὸν οἰδὲ \(\deltaή\) τὸ ἀναγκαῖον, οἰδὲ \(\δή\) ὅλως \(\text{ἡ ἔνος ἀνδρός}, \) μὴ βασιλεῖως ὅτις \(\text{ἡ τινος} \) τοιοῦτον· \(\text{ὁ \(\text{δὲ} \) νόμος} \) ἀναγκαστικὴν ἔχει δύναμιν, \(\text{λόγος} \) ὅν \(\text{ἀπὸ τινος} \) φρονήσεως καὶ νοῦ. \(\text{καὶ τῶν} \) \(\text{μὲν} \) ἀνθρώπων \(\text{ἐχθαίρουσι} \) τοὺς \(\text{ἐναντιουμένους} \) ταῖς ὀρμαῖς, \(\kappaἀν ὁρθῶς \) αὐτὸ δρῶσιν· \(\text{ὁ \(\text{δὲ} \) νόμος οὐκ} \) \(\text{ἐστι} \) ἐπαχθῆς \(\text{τάττων} \) τὸ ἐπιεικές.\)

---

\(^{81}\) ENI.2 1094a21–b11.

\(^{82}\) οἱ γὰρ νομοθέται τοὺς πολέτας ἐδίκαιος ποιοῦν ἀγαθοῖς, καὶ τὸ μὲν βούλημα παντῶς νομοθέτου τοῦτ’ ἐστίν, \(\text{ὁσοι \(\text{δὲ} \) \(\text{μὴ} \) \(\text{ἐφ} \) αὐτὸ ποιοῦν ἁμαρτάνουσιν, καὶ \(\text{διαφέρει} \) τούτῳ πολιτεία πολιτείας ἀγαθῆ φαύλης.} \) (ENII.1 1103b3–6)
Now, the command of a father is not strong or compelling—the command of a single man certainly not at all, unless he is a king or someone of that sort—. But the law [nomos] has a compelling power, since it is logos that originates in a certain phronēsis and intellect. Also, [people] hate those among men who oppose their impulses, even if they are right to do so; but the law is not grievous when it orders what is reasonable. (ENX.9 1180a18–24)

Proper habituation will necessarily face resistance, because it implies making people go against their innate tendencies toward pleasure and laziness (§1.3.2D). So a strong, compelling power is needed to do the job. Individuals often do not have the sufficient strength to get it done, but the law does, because it is a logos. This link between logos and the laws is puzzling and should be explained. For now, however, I want to point out that this passage reveals that for Aristotle care for habits should take place first and foremost at the political level. In the following lines, he reproaches the fact that among all cities only in Sparta and few others “the legislator seems to have cared [epimeleian pepoiēsthai] for upbringing and practices”, while the other cities are absolutely careless [exēmelētai] about it, and the task is left to each man, who legislates as he pleases over his wife and children, just like the Cyclopes used to do. From this Aristotle concludes:

Therefore, the best thing is for there to come to be a correct, public care [koinē epimeleian] [...] For it is clear that public care is generated from laws, and that the most convenient public care is the one that derives from excellent laws. And it does not seem to make a difference whether the laws are written or unwritten, or whether one or many will be educated through them, just as it does not in music, physical exercise, or the other practices. (ENX.9 1180a29–b3)

83 “It is difficult for people to come across the right education towards virtue from childhood if they have not been brought up by adequate laws, because the life of moderation and perseverance is not pleasant for the many, particularly for the young. [εἰ νῦν δ’ ἀγωγὴς ἁρμῆς τυγχαίνει πρὸς ἀρετήν χαλεπὸν μὴ ὑπὸ τοιοῦτος τραφέντα νόμους· τὸ γὰρ σωφρόνως καὶ καρτερικῶς ζῆν ὡχὶ ἀδύνατον τοῖς πολλοῖς, ἄλλως τε καὶ νέοις.] (ENX.9 1179b31–34).

84 ἐν μόνῃ δὲ τῇ Δακεδαιμονίῳ πόλει μετ’ ἄλλων ὁ νομοθέτης ἐπιμέλειαν δοκεῖ πεποίησθαι τροφής τε καὶ ἐπιτηδευμάτων· ἐν δὲ ταῖς πλείσταις τῶν πόλεως ἐξειλήγη περὶ τῶν τοιούτων, καὶ ζῇ ἕκαστος ὡς βούλεται, κυκλοσκικῶς θεμιστεύων παιδῶν ἀλόχου. (1180a24–29). About the Cyclopes, Homer says:

they do not hold deliberative assemblies or obey received laws; they dwell in hollow caves on the summits of high mountains, where each lays down the law to his own children and wife, not associating with one another [εἰδ’ όλλων ἀλόχου] (Odyssey [Od.] IX 112–115; tr. Depew).
Consequently, Aristotle’s depiction of the best constitution includes public care institutions overseeing the rearing and education of the young. Areas of public care he discusses include when and with whom people should marry and reproduce (Pol. VII.16), what physical exercises children of different ages should practice (VII.17), and what kinds of stories, speeches, and even musical harmonies the young should be exposed to VIII.2, 5–7). His notion of public care is thus an amalgam of what we would call ‘education’ and ‘paternalistic intrusions into individual privacy and freedom’.

This passage leaves two important questions open: first, in what sense is the law a logos? And second, how does Aristotle justify his radically paternalistic conception of public care? Let us start with the former, since answering it will help in answering the latter.

**Law and reason**

A deflationary reading of the claim that laws are a logos would interpret ‘logos’ in the sense of ‘rational speech’: laws are, in fact, discourses that proceed from people’s exercises of rational thought (phronēsis in the best cases). But other things that Aristotle has to say about the law suggest that this reading is too deflationary—or, to be precise, too individualistic. In the context of Aristotle’s political thought, laws are not just speeches, but a kind of rational structure whose function is to preserve the form of the semi-organism that is the polis. Explaining this requires briefly mentioning some of the basic aspects of Aristotle’s political thought.

First, the polis is something that exists by nature, and has a natural function. Because Aristotle’s polis is a hylomorphic semi-organism, he feels free to explain the political world by reference to the same conceptual tools he uses in biology. Thus, like any matter-form composite, the polis’ identity is defined by its form. Because the city’s identity changes when its constitution (politeia) changes, it is clear that the constitution is the form of the political community. Its territory and individual inhabitants, though necessary, can change without the city’s identity being altered, so they are rather conceived as equivalent to its matter.

The constitution, in turn, is portrayed as a city’s structure (taxis). The different constitutions (e.g. monarchy, aristocracy, democracy) specify the way in which a city is to be

---

85 The polis is natural: Pol. I.2 (1252b27–34; 1253a1–2). The polis has a function: Pol. VII.4 1326a13–14.
86 it is unclear whether these methodological tools themselves allow for an application beyond the level of the organism. Johnson (2005, 237–246) argues that they do not. So Aristotle’s naturalization of many oppressive relationships may have resulted from his misappropriation of his own methodologies.
87 Pol. III.3 (1276a17–1276b11).
88 Pol. III.6 (1278b 8–9); IV.1 (1289a15–20); IV.3 (1290a7–10).
governed, by establishing how government is divided into offices, who is able to enter those offices, how rulers and ruled are related, and the inhabitants are to be educated. Also, importantly, each constitution defines the city’s goal (i.e. its guiding conception of the happy life). Because of this it defines the way of life that characterizes the city, and the overall character that corresponds to the people living in it. Thus, in the hylomorphic interpretation of the polis, the constitution would be equivalent to its formal and final principles.

For a community to have a constitution, it must be ruled in accordance with appropriate laws. Laws are mandates that establish universal obligations (as opposed to the particular obligations established by decrees, or the whims of a dictator). If a community’s laws (that e.g. specify the tasks of public offices and determine who should be rulers and who should be ruled) are not followed, it is clear that the constitution is not really the community’s structure—so arguably wherever this happens the community does not really have the form of a polis. So for constitutions to function particular instances of rule must be informed by universal laws. Moreover, Aristotle subordinates the laws to the constitution: “all laws must be established with respect to the constitutions, but not constitutions with respect to the laws”. So the purpose of laws is to establish generalizable practices of rule that serve as appropriate, homogeneously applicable, and consistent means to reach the city’s goal, determined by its constitution.

This brings us back to our main concern: laws, being the product of phronēsis and intellect, are a kind of logos. They must establish proper generalizable practices to enact the city’s constitution and its goals. But they are a logos not only in the sense of an argument or a discursive statement produced by intellect: they also are a logos in the sense that they are intellect itself. This is made explicit in a discussion about absolute kingship, a form of government in which the king’s will is the rule (Pol. III.16). Aristotle holds that it is more choiceworthy to have the laws rule rather than individual people’s wills. Of course, particular people are needed to execute the laws, but they should be understood as “guardians of the laws and assistants of the

---

89 The best constitution, for instance, structures the city in a way that anyone of its citizens may be happy (Pol. VII. 2 1324a23–25). For the polity’s goal, see IV.11; for democracy’s goal, e.g. V.9.

90 For the constitution as the polis’ way of life [bios], see Pol. IV.11 1295a40–b1. For its role as defining the citizens’ character, see Pol. VIII.1 1337a14–18.

91 Pol. IV.4 discusses populist democracies (in which the multitude erects popular leaders who rule on the basis of case-by-case decrees [psēphismata] rather than universal laws [nomai]). Aristotle ultimately denies that this kind of community has a constitution at all, because “wherever laws do not rule, there is no constitution. For it is necessary for the law to govern all universal cases [ὅπου γὰρ μὴ νόμοι ἀρχοῦσιν, οὐκ ἔστι πολιτεία. δεῖ γὰρ τῶν μὲν νόμων ἀρχεῖν πάντων τῶν καθόλου]” (1292a32–33).

92 πρὸς γὰρ τὸς πολιτείας τοὺς νόμους δεῖ τίθεσθαι καὶ τίθενται πάντες, ἀλλ’ οὐ τὸς πολιτείας πρὸς τοὺς νόμους. (IV.1 1289a13–15)
laws” rather than sources of rule. The main reason for preferring laws to particular humans is that the latter introduce corruption into the system, by making particular decisions that aim at e.g. favouring their friends and advancing their own careers, instead of aiming at the universal order that preserves the constitution.

ὁ μὲν οὖν τὸν νόμον κελεύων ἄρχειν δοκεῖ κελεύειν ἄρχειν τὸν θεὸν καὶ τὸν νόμον μόνου, ὁ δ’ ἄνθρωπον κελεύων προστίθησαι καὶ θηρίον; ἢ τε γὰρ ἐπιθυμία τοιούτου, καὶ ὁ θυμὸς ἀρχοντας διαστρέφει καὶ τοὺς ἀρίστους ἄνδρας, διόπερ ἀνεύ ὀρέξεως νούς ὁ νόμος ἔστιν.

Thus, he who urges for the law to rule seems to urge for God and nous alone to rule, but he who urges a man to rule also assents to the rule of a beast; for appetite is of such kind [i.e. beastly], and spiritedness, when in power, perverts even the best men. The law, therefore, is intellect without desire. (1287a28–30)

Aristotle’s statements that the law is logos and intellect should be understood in the light of his claims that the polis can be theoretically modelled as a semi-organic, hylomorphic entity whose form is the constitution, and that the constitution effectively structures the polis through the application of universal laws. So, just as the rational soul is the form of the human composite, the constitution and the laws should be understood as the soul-like aspect of the semi-organic political community. If this is so, to carry the analogy forward, just as our eyes, heart, limbs, and other organs are the matter that allows us to actualize our soul’s capacities, in the same way the individual rational human beings that serve as “guardians and assistants of the laws” are the material substratum allowing for the polis’ rational self-maintenance and self-governing activities. Thus, given his metaphysics of the polis, when Aristotle calls the laws ‘reason’ and ‘intellect’, he probably means it not just in the minimal sense of a rational and intellectual product of some individuals’ reasoning, but also in a stronger sense of a rational semi-psychic form that emerges from these individuals’ activities.

In fact, as long as the organicist logic applies, the private logos of individuals must be understood as the material tools through which the public logos of the polis actualizes its existence. These views of the relationship between individual and polis follow from Aristotle’s strong metaphysical claims about the priority of the polis with respect to individual citizens and their cognitive capacities.

93 κἂν εἶ τις ἄρχειν βέλτιον, τούτους καταστατέαν νομοφύλακας καὶ υπηρέτας τοῖς νόμοις (1287a20–22).
Public vs. private reason

In fact, according to Aristotle, just as the whole explains the nature of the parts in other cases, the nature of the polis should also explain the nature of its constituents—villages, households, elementary relationships (master-slave, husband-wife, father-child), and individuals. This methodological principle reveals the extent to which he is willing to subordinate individual agents, both politically and metaphysically, to their political wholes. This implies, of course, rejecting the idea that individuals fully own themselves, not only in the case of slaves, women, and children (groups that Aristotle explicitly subordinates to the free adult male’s rule), but even in the case of free citizens:

οὐδὲ χρὴ νομίζειν αὐτὸν αὐτοῦ τινα εἶναι τῶν πολιτῶν, ἄλλα πάντας τῆς πόλεως, μόριον γὰρ ἐκαστὸς τῆς πόλεως.

One should not consider that any of the citizens belongs to himself alone, but rather that they all belong to the polis, since each one of them is a part of the polis (Pol. VIII.1 1337a27-29)

This is nothing but the ultimate consequence of Aristotle’s view that humans are a hyper-political species. Political species are those that have a common function that they must perform by cooperation. It is because of this that humans are rational in the first place: if nature gave humans logos it must be explained, like any other organic trait, on account of their biological needs. The role of logos in human life is allowing for the communication of the just and the unjust, so that complex political cooperation can be achieved to such a level that the existence of the polis is possible.

So humans have logos because they lead a hyper-political existence; and they lead a hyper-political existence because otherwise their own good (a life of excellent rational activity) would be unattainable. In fact, individual humans are not self-sufficient: they must rely on their common activities in order to survive, and a fortiori in order to achieve a good life. Of course, individual humans are necessary to constitute a polis in the first place (and its inventor is to be praised for having done the greatest possible good to everyone else), but although these

---

94 Pol. I.1 1252a17–23.
95 HA I.1 (488a9).
96 Pol. I.2 1253a7–18.
97 See the discussion of self-sufficiency [autarkeia] in EN I.7 (1097b6–16). In the Politics, Aristotle argues that the kind of self-sufficiency that allows for subsistence can be attained at the level of a household, but the ultimate self-sufficiency that allows for us to lead a good life of rational activity is achievable only in a city (Pol. I.2 1252b12–14; 27–34).
individuals must have been temporally prior to the polis, the ontological order demands that the
polis be prior in explanation and in being: it is by reference to the nature of the whole that we can
understand the nature of the parts.

καὶ πρῶτον δὲ τῇ φύσει πόλις ὡς οἰκία καὶ ἕκαστος ἡμῶν ἔστιν. τὸ γὰρ ὅλον πρῶτερον ἀναγκαῖον εἶναι τοῦ μέρους· ἀναιρουμένον γὰρ τὸ ὅλον οὐκ ἔσται ποὺς οὐδὲ χεῖρ, εἰ μὴ ὀμονόματι, ὀστέρος εἰ τις λέγει τὴν λαθήν (διαφθαρεῖσα γὰρ ἔσται τοιαύτη), πάντα δὲ τῷ ἔργῳ ὁρίζεται καὶ τῇ δυνάμει, ὡστε μηκέτι τῷ οὐκέταν τὰ αὐτὰ ἔσται ἀλλὰ ὀμονόματα. ὅτι μὲν οὖν ἡ πόλις καὶ φύσει καὶ πρῶτερον ἡ ἕκαστος, δῆλον.

The polis is also naturally prior to the household and to each one of us. For the whole is
necessarily prior to the part, because if the whole is destroyed there won't be a foot or a hand,
unless one speaks homonymously, as if someone called ‘hand’ one made of stone (for the dead
hand would be like that). And all things are defined by their function and their capacity, so
that we should not say such things are the same, but that they are homonyms. It is thus clear
that the polis is by nature, and is prior to each individual. (Pol. I.2 1253a18–26)

The analogy in this passage is quite explicit: just as a hand cannot perform its natural function if
separated from the rest of the body, likewise the human individual cannot perform its natural
function if separated from the polis. The human function is rational action, so the analogy
implies that individual rationality depends on participation in the polis. Individual reason is thus
a political achievement: it flourishes on the fertile soil of the political communities, but shrivels
if separated from it. So if we were cut off from society and left to our own devices early on—first
of all, we would probably die; but even if we did not—we would not be likely to develop the
active dispositions that are necessary to truly engage in a life of fine rational activities. Our
cognitive capacities would be likely to remain close to the animal level, unable to exert any
significant amount of effort to go beyond the search for pleasure linked with the immediate
satisfaction of basic biological needs. This is why Aristotle claims that a human being who lives
in isolation is either a beastly sub-human or a divine super-human. Because an individual
separated from his polis cannot be rational.

To bring it all together, individual reason depends on the collective achievement of a
public rationality that is embodied in good laws and a good constitution, and enacted by the
collective enterprise of excellent public administration that sustains the existence of the polis.
Without good laws, a prosperous city, and correct public care (which habituates people into the

99 Pol. I.2 1253a1–5; 1253a26–29.
character dispositions that correspond to the polis’ constitution), we would be unable to lead rational lives. It thus must be concluded that the rationality of the part (i.e. the individual rational agent) is explicable only by reference to the rationality of the whole (i.e. the laws).

This leads to the section’s last question: why is habit care a part of public administration?

Why care should be public

μέγιστον δὲ πάντων τῶν εἰρημένων πρὸς τὸ διαμένειν τὰς πολιτείας, οὐ νῦν ὀλεγωρούσι πάντες, τὸ παιδεύεσθαι πρὸς τὰς πολιτείας. οὐφελοσ γὰρ οὔθεν τῶν ὠφελιμωτάτων νόμων καὶ συνδεδοξειμένων ύπὸ πάντων τῶν πολιτευομένων, εἰ μὴ ἔσονται εἰθαμένοι καὶ πεπαιδευμένοι ἐν τῇ πολιτείᾳ.

The most important of all the mentioned ways to make a constitution stand firm, which everyone neglects nowadays, is to be educated towards the constitution. For the most beneficial laws are of no avail, even if all the citizens jointly approve of them, if they will not be habituated and educated with respect to the constitution […]. (Pol. V.9 1310a12–19)

Even the best constitution with the best laws, endorsed by the population, will not last long if people’s character is such that they do not follow the law. Aristotle calls this a case of akrasia at the level of the polis (1310a18–19). So, to avoid the constitution’s disintegration due to political akrasia, the whole population must be habituated in accordance with it.

Aristotle expands this argument later on in the Politics (VIII.1). After recalling that constitutions are harmed wherever the legislators do not pay attention to the education of citizens’ character, he justifies the requirement that habit care should be public by reference to the (already cited) part-whole relationship between individual and polis, and the polis’ natural teleology:

ἐπεὶ δ’ ἐν τῷ τέλος τῇ πόλει πάση, φανερῶν ὅτι καὶ τὴν παιδείαν μίαν καὶ τὴν αὐτὴν ἀναγκαίον εἶναι πάντων, καὶ ταύτης τὴν ἐπιμέλειαν εἶναι κοινῇ καὶ μὴ κατ’ ἰδίαν, ὅν τρόπον νῦν ἐκάστος ἐπιμελεῖται τῶν αὐτοῦ τέκνων ἰδια [...]. ἀμα δὲ οὐδὲ χρὴ νομίζειν αὐτὸν ἀυτοῦ ταῦτα εἶναι τῶν πολετῶν, ἀλλὰ πάντως τῆς πόλεως, μόριον γὰρ ἐκάστος τῆς πόλεως: ἣ δ’ ἐπιμέλεια πέφυκεν ἐκάστου μορίου βλέπειν πρὸς τὴν τοῦ ὅλου ἐπιμέλειαν.

Since the the whole polis has a single goal, it is clear that the education must be one and the same for everyone, and care for it should be public, and not private, in the way that currently each person privately cares for his own children […]. At the same time, one should not consider that any of the citizens belongs to himself alone, but rather that they all
belong to the polis, since each one of them is a part of the polis; and it is natural that the care of the part is with a view to the care of the whole. (Pol. VIII.1 1337a27-29)

Public care’s role is thus habituating the individuals so that the city as a whole reaches its common, natural goal. Although this implies the full subordination of individual education to the goals of the city as a whole, the anti-individualistic aspect of this view could be overstated. The goal of the polis, after all, is that its citizens lead a happy life, and this is possible only within a rationally structured polis. So individual habituation must be organized publicly because this is the only way to make individuals themselves able to develop the rational dispositions of character that will later on allow them to take their habituation into their own hands, and develop it in accordance with their own reason. It is because agents can take habituation into their own hands that, in EN III.5, Aristotle can point to the careless individual and say that he could have cared for his habits, so he is responsible for his character and his actions. We are joint causes of our own character and phantasia of the good insofar as we can engage in self-habituation; but self-habituation is possible only if we have received appropriate public care. Thus, from Aristotle’s perspective, centralized public habituation is the only way in which individual autonomy is possible—to the extent that it is possible.

To sum up, in this section I have argued that (1.4.1) care for our character is a necessary condition for human agentive control, and that (1.4.2) Aristotle understands this care as a primordially political endeavour. The consequence of these two lines of reasoning is that we are able to exercise rational control over our lives as individual agents only within the proper social and political structures that make it possible for our rational capacities to be properly extended through habituation.

1.4.3. Indirect Intellectualism

It is time to go back to the Chapter’s main question, and to bring the pieces of the answer together. The Chapter started with the question of human agency: what is it about human adults that makes us agents in a more complex sense than animals or children? After (§1.1) introducing Aristotle’s vocabulary of action and sketching the main principles of animal agency, I argued that (§1.2) the available options for an answer to the question are not fully satisfactory. On the one hand, intellectualism cannot account for complex human agentive control, because the

---

100 For an in-depth analysis of this and other arguments Aristotle gives in favour of public education, see Curren (2000, esp. Chapter 5).
intellectualist move that we can determine the goals of our actions by reasoning about them is blocked if we take seriously (as we should) the copious textual evidence for the division of labour between character and reason, according to which the former sets the goal and starting point, and the latter sets the things toward the goal. But anti-intellectualism is not satisfactory either, because it cannot satisfactorily explain how reason is supposed to guide and rule over desire if the former cannot shape the latter. Moreover, if we cannot set our goals through reason, Aristotle’s ethical enterprise would be pointless, since it is a rational reflection about the goals of human life.

Through Aristotle’s clue that we can control our goals (i.e. our phantasiai of the good) by controlling our character, I set out to explore (§1.3) how the process of character development, or habituation, works. I found that habituation is a process whereby we interiorize external principles of motion by means of repeatedly practicing actions of the same kind. This practice extends our natural capacities because it makes our active and receptive dispositions able to endure more difficult and longer periods of activity. This resistance buildup is necessary to resist temptations, pursue complex long-term goals, and perform demanding actions, like those of theoretical reasoning and deliberation. If that is habituation, we can understand moral and intellectual development as the acquisition of the active and receptive dispositions that allow us to be properly affected by objects (i.e. experience the proper pleasures and pains towards them), and properly engage in noble rational actions, at the same time feeling pleasure in performing them.

This exploration of the nature of habituation revealed that the key structural difference between animal and human agency is that the latter has a circular structure in which actions, character, the appearance of the good, desire, and reason mutually shape each other. This opens the path to answering the question about human agency: we can rationally set our life’s goals (our phantasiai of the good) by carefully shaping our character through habituation.

Now, the ‘we’ in the previous paragraph turns out to be very important, since (§1.4) habit care is for Aristotle both a necessary condition for individual human agency and a primarily political endeavour. Individual agents can engage in rational action control and rational habit-care only if they have received from others the proper early habituation that develops in them the necessary resistances to direct their actions in accordance with reason—not an easy thing! Thus rational care for character is a political task. I traced the notion of care [epimeleia] from Aristotle’s discussion of action through his political theory, and concluded that the individual rational dispositions that allow us to control our actions depend on a fertile soil of public rationality, expressed by the polis’ laws, without which individual humans could not lead rational lives. The polis is prior to the individuals as the whole is prior to the parts. In the same way, public rationality is prior to private reason.
Where does all this leave us? I want to conclude by defending two views: first, reason can shape our goals indirectly, via shaping character. Second, individual rational control is an extension of our natural control capacities, which is possible only thanks to the public rationality embodied in social practices of habituation and legislation.

**Indirect Intellectualism, the role of ethics, and the human–animal distinction**

(A) *The two theses of indirect Intellectualism*

There are two main ways in which an individual’s rational control over her actions is indirect. Here is the first one:

*Indirect Intellectualism 1:* Reason determines the goals of action via careful shaping of character.

This thesis has the advantage that it can accommodate the key insights from both direct Intellectualism and anti-Intellectualism: it preserves the division of labour (it is character, not reason, that directly determines our phantasia of the good), but it also explains how the soul’s rational part can guide and rule over the non-rational part.

The functions of reason discussed above (specificatory, instrumental, and coordinating) are precisely what individual agents require to shape their habituation patterns. Just as an athlete training for a competition must rationally select appropriate actions, and coordinate them toward doing well at the time of the competition, the autonomous individual agent must rationally select the specific practices, and coordinate them through time toward the acquisition of the character she wants to build for herself.

But how does she know what character she should build for herself in the first place? This is the terrain in which the anti-Intellectualist is still right: the originary appearance of a certain character as good must necessarily precede all reasoning and deliberation about habit-care. The first phantasia of the good character is a consequence of the first steps of an individual’s habituation, which did not depend on her, but were performed by other people who cared for her. This leads to the second way in which rational control over actions is indirect:

*Indirect Intellectualism 2:* Public reason allows for individual reason’s control over goals by extending the natural capacities of individuals through public care.
An individual’s ability to rationally set her own goals requires that her innate tendencies toward excessive pleasure and away from hard work were corrected early on in life. But we cannot do this to ourselves: it is a task for the community around us, including our parents in so far as they embody public values. So reason can be in control of goals only via public reason’s early public care. This way, it may be said that public reason indirectly allows for individual rational control over goals by developing in the individual agent her capacity for rational self-habituation.

(B) What the study of ethics is for

Now we can see the point of studying ethics. As mentioned earlier, the study of the human good belongs to political science. This entails that the theory of the human function, the difference between the ethical and theoretical virtues, the study of friendship and pleasure—they are all aimed ultimately at improving the political community. This implies improving the lives of individual humans, since the goal of the community is that its citizens lead happy lives. But from the political perspective an individual citizen’s happiness can be seen as a means to achieving the political goal of a flourishing city. So nothing stops reason from deliberating about the human good, insofar as it is part of the things toward the (political) goal.

Aristotle calls political science architektonikê: master-builder. Elsewhere he says that one of the kinds of phronēsis is also architektonikê. Crucially, phronēsis in its architectonic sense is identical to the science of legislation. So the statesman’s legislative virtue is phronēsis itself, insofar as it is in charge of orchestrating the city’s development. The legislator’s task thus includes deliberating about the individual human being’s good as a means toward the polis’ goal, and writing laws that advance that goal. As seen above, such legislative task to a great extent consists in establishing the proper mechanisms of public care for early habituation.

If we see the Ethics and the Politics in this way, as continuous parts of Aristotle’s contribution to political science, it becomes clear what the role of ethical study is in becoming better: it is an intellectual reflection about the human good that is ultimately aimed at developing good cities. A large part of this endeavour would consist in understanding how human individuals must be habituated so that they can afterward lead a life of self-care and rational control over their actions.

101 EN1.2 (1094a27), VII.11 (1152b2).
102 EN VI.8 (1141b21–26).
103 For congenial readings, see e.g. Kraut (2002, 92–97) and Frede (2013, 32–33).
(C) The difference between animals and humans

Recall that there are two senses of the word ‘intellectualism’ (§1.2.1): Intellectualism (capitalized ‘I’) is the view that reason can set goals; and intellectualism (lowercase ‘i’) is the view that the structural difference between animal and human agency is the latter’s possession of reason. I just argued that an indirect version of Intellectualism is the better interpretation of Aristotle’s theory of action. But now I want to say that when it comes to the distinction between animal and human agency intellectualism does not tell the whole story. A fundamental difference between human and animal agency that is not captured by intellectualism is that the former has a circular structure and the latter has a linear one. I should clarify, at any rate, that the distinction cannot be clear-cut, because some animals participate in habituation to some minimal extent, which implies that their appearances of the good can to some minimal extent be shaped by their past actions. Thus some animals—arguably the most intelligent ones, since habituation implies understanding significations—occupy a middle ground between the human circularity and animal linearity. Anyhow, a mention of the thorough-going human circularity is necessary to explain the greater control over actions that human agency allows for, and this is just not part of the intellectualist story.

But reason, on the other hand, is necessary for the careful habituation that circular agency makes possible. So a complete characterization of human agency must refer to the powers of reason. But importantly, this complete characterization should actually refer to two powers of reason: private and public, since public reason is indispensable for the individual to become capable of rational self-habituation, and the latter is necessary for individual control over actions (see Fig. 1.5).

---

104 This is not to say that some animals are self-habituating. Animals need someone else to train them; the capacity to direct one’s own habituation is an exclusively human trait.
To conclude, what makes human agency more complex and thorough than animal agency is a mix of (1) its circular structure, (2) the rational public care that is possible within the political community, and the individual self-habituation that private reason, together with 1 and 2, make possible.
Mind is a leaky organ, forever escaping its ‘natural’ confines and mingling shamelessly with body and with world. [...] our behavior is often sculpted and sequenced by a special class of complex external structures: the linguistic and cultural artifacts that structure modern life.

—A. Clark: *Being there*

Throughout this Chapter, I have talked about the upshot of public care and private self-habituation in terms of the *extension* of our natural abilities. With this I mean to express how we (1) enlarge our innate dispositions by making them more resistant through practice, and (2) become able to constrain the direction of this expansion through our expanded rational capacities. I chose to talk of extension to bring forth resonances between Aristotle and the extended-mind theory in contemporary philosophy of mind (§2.1.2F). I think the parallels between the two theories are worth exploring. I will do some of this exploration later on, but for now I want only to stress the following.

The extended-mind theory holds that our capacities for cognition and action depend on the fact that our cognitive processes are not skull-bound, but rather include parts of our environment that actively contribute to these processes. Without such extensions of our cognition into our environments, we would be unable to perform the complex action control that we can do. So agentive control is dependent on our being properly linked with an environment that extends our natural capacities.

Similarly (though by no means identically), Aristotle’s theory as interpreted here states that our capacities for cognition and action depend on the habituation process that extends them by interiorizing principles of action that were initially external to us. This is a corollary of indirect intellectualism: individual rational control is an extension of our natural control capacities, made possible only by the social structures that surround us. If this is true, there is an interesting parallel between the two theories that is worth exploring.
§1.5. Looking ahead

I want to close this Chapter by highlighting how a discussion of Aristotle’s views can shed light on our own philosophical concerns about action.

First, there are many elements of his account that simply do not work, are unjustified, and therefore not relevant (e.g. his naturalization of relationships of domination between human individuals, the view that some people are by nature owners of their own selves, the claim that those dedicated to certain jobs cannot participate in politics, and the idea that individuals must be subordinated to the polis as an organic part is subordinated to the whole organism). But Aristotle’s main insight—i.e. that our ability to rationally control our actions depends on the way in which our public environment is structured—is as valid today as it was in his time.

Explaining his approach to the study of ancient Greek philosophy, John McDowell says his aim is “to counteract a way in which […] modern prejudices about rationality tend to distort our understanding of Greek ethics” (1998b, vii). I am sympathetic to this general approach: one of the benefits of rigorous study of thinkers from another time and culture is that it opens an opportunity for us to become aware of our own prejudices as such. My hope is that the present engagement with Aristotle’s theory of action may serve to bring to light a modern prejudice about rationality that escaped McDowell’s own interpretive efforts. I am talking about the prejudice of excessive individualism that tends to distort our understanding of Greek ethics, and also of ourselves as agents.

A large portion of contemporary exegesis (and perhaps also of contemporary philosophy) tends to conceive of rationality as a faculty that belongs to an individual mind, and of the individual as the almost self-sufficient source of rational activity. If my interpretation is right, Aristotle did not share with us moderns this individualistic prejudice; he rather conceived of individual reason as dependent on common resources of public rationality. If disconnected from this resource, individuals are incapable of rational action. This shows that reason is not a property they have merely by their own individual nature.

I think something like this is turns out to be the case. Our cognitive capacities are so feeble and malleable that we can successfully engage in rational thought only in very specific circumstances and environments. Modern life has made it a lot easier for those circumstances to take place (albeit bringing new obstacles of its own), but reflection still is a fragile thing, and rational control over action is a path full of obstacles. Or so I will argue in the following Chapters: our current empirical understanding of human cognition and action casts serious doubt upon the modern view of human individuals as largely self-sufficient, rational decision-
makers. Our rationality is, in fact, much more limited than a modern thinker like Descartes or Hobbes (or in fact any of us in a non-reflective attitude) would portray it to be.

If this is so, we should pay more attention to the way our environment is set up in order to protect and extend our limited capacities for reflective control over action. This is crucial because individual autonomy depends on it.

From this perspective, Aristotle’s theory that our rational control is based on a proper habituation into practices that extend our rationality by making us more resistant to effort gains a novel relevance. But I will argue that most of the extensions we need in order to lead autonomous, rational lives, are of a very different kind. No amount of cognitive training will suffice in many cases, because our natural reflective capacities are just too limited, and our practical world is just too complex, for us to reach the level of a fully developed virtuous person, who is able to remain unaffected by the situation’s irrelevant details, effectively grasp its relevant features, and act accordingly. In contrast, our decisions and actions are surprisingly often framed in ways that leave out relevant information, and are influenced by irrelevant information. I will argue (in Chapter 5) that these framing effects are unavoidable, and that because of this cognitive training is not a sufficient tool to generate and protect individual autonomous control. To deal with the obstacles to autonomy, we need to think about other forms of rationality extension, beyond what habituation practices can offer. These go in the direction of institutional and environmental design.

For now, however, the next Chapter will jump to times closer to us, and present a landscape of recent theoretical positions about the notion of intentional action, and the subsequent conceptions of agentive control that stem from them. This will allow us to bring the discussion of agentive control into the contemporary terrain.
Chapter 2: What is control?

A theoretical landscape of intentional action

This work’s ultimate goal is to advance towards a framework of human action that is useful to understanding human agency at a social level, and takes our cognitive limitations into account. Such a framework should be useful in finding answers to questions like: What is the nature and reach of autonomy in everyday social life? How, and to what extent, am I in control of my behaviour, and how, and how much, is it controlled by my external environment, or by blind, unintentional cognitive processes within me? Is it right to say that my actions are often a product of my “conditions of existence” (things like my social class, my parents’ profession, my cultural environment) more than my self-chosen beliefs, motivations, and commitments? If we are to some extent self-determined agents, then how does self-determination work, what are its limits, and what can we do to bolster it and protect it from external and internal threats?

Answering such large questions, in accordance with what we know empirically about human cognition and action, is too large a project for a single dissertation. What the following Chapters attempt to do is, rather, investigate the basic issues and concepts in the psychology and philosophy of mind and action that would ground a proper answer to any of them. This attempt will lead to an examination of three interrelated notions: agency, automaticity, and control, and the problems involved in discussing them. The guiding question is: what kind of agency and control do we humans have over our behaviour, if most of it is automatic? How does it work, and what are its limits? These are, I think, the fundamental questions that need to be rigorously addressed before attempting to answer any of the questions above.

The present chapter presents the philosophical landscape of theories concerning intentional action, within which the exploration of these issues will unfold. First, (§2.1) I describe the two main philosophical approaches to the explanation of intentional action: intellectualism and anti-intellectualism, and the two notions of intentionality that correspond to each one: operative and act intentionality. Broadly, intellectualism is the view that action production requires the rule-based manipulation of mental states by a central executive capacity; and anti-intellectualism opposes this by exploring the ways in which non-reflective, skilful, bodily interaction with the practical environment sufficiently counts as action.

The advent of connectionist or association-based models of the mind, together with evidence in favour of embodied and embedded explanations of the mind seemed to give the
upper hand to anti-intellectualism. But in §2.2 I describe how intellectualism has turned out to be quite a supple framework, able to shift shapes by taking on board its opponent’s ideas about embodiedness and situatedness. I thus narrate the birth of situated intellectualism through the incorporation of the associative neural-networks architecture and the broadening of concepts like intention and planning.

As it currently stands, the debate (§2.3) is between the situated intellectualists who think behaviour without reflection is nothing but automaticity and reflex, and anti-intellectualists who consider reflection to be too blunt a tool to account for the fine-grained control we display in skilled action. While the situated intellectualist claims to be able to explain how reflection is necessary for skilful action, the anti-intellectualist argues higher-order reflective processes are unnecessary, even harmful, for the performance of such actions.

Figure 2.1: A theoretical landscape of intentional action.
§2.1. Intellectualism and anti-intellectualism about intentional action

What makes a behaviour an intentional action, instead of, say, a mere bodily motion? The very way of framing this question (common at least since Wittgenstein)\(^1\) suggests that intentional actions are mental phenomena: this is what distinguishes, say, a wink from an ocular twitch, and allows us to refer to actions as ‘intentional’ in the first place. In fact, it is common to consider intentionality a trademark of the mental. But what does the word ‘intentional’ mean? In re-introducing the scholastic notion of intentionality into modern philosophical discussion, Brentano employed two metaphors: intentional states or acts are understood as directed towards something, and being of or about something (i.e. having something as its content).\(^2\) Often the two metaphors (directedness and aboutness) are interpreted as synonymous, but this is not necessarily so (cf. Reuter 1999). In fact, the two main perspectives on intentional action can be seen as interpreting intentionality by emphasizing either of these metaphors over the other.

2.1.1. Intellectualism

On the one hand, there is the view that for a behaviour to be an intentional action its production must be explained by reference to mental states that are about something. The fact that, in order to explain a certain behaviour, we must refer to such mental states as beliefs, desires, and intentions, justifies our calling said behaviour an intentional action. That is because said mental states “contain an object intentionally within themselves” (Brentano 1874/2009, 68), and this object, represented in the agent’s mind (as a goal to be achieved or an intention to be carried out), establishes the behaviour’s conditions of satisfaction or failure. The behaviour thus acquires the normative flavour characteristic of intentional action (it is “rationalized”) when placed in relation to the representational content of the mental states that account for its performance.

Thus, from this perspective, a behaviour counts as an intentional action insofar as its production requires that (1) the agent entertains the mental representation of a goal to be achieved (along with other representations about relevant features of the world, and what must be done in order to achieve the goal); and (2) the agent manipulates those inner representations in order to

---

\(^1\) “[W]hat is left over if I subtract the fact that my arm rises from the fact that I raise my arm?” (Philosophical Investigations §621.)

\(^2\) “Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object” (Brentano 1874/2009, 68). In this context, the words ‘content,’ ‘object,’ and ‘something’ should be understood, not as meaning ‘some objective thing,’ but rather in their broadest possible sense, as anything, either material or not, existent or not, that is different from the intentional state itself and that the intentional state can therefore intend.
produce a behavioural output that is structured in order to attain the represented goal. Condition 1 makes this a *representationalist* approach, since it requires the existence of inner representations; condition 2 makes it an *intellectualist* approach, since it requires that an intellect manipulate these representations and coordinate behavioural operations in accordance with them.

What do I mean by ‘representation’? In a broad and schematic understanding, a representation is the content of a mental state that serves as a symbol for something external to the mental state itself (which could be, for instance, a feature of the world or another mental state). Symbols, however, like words on a page, are mute without interpreters who perform some intellectual operation of interpretation on the symbols in order to ‘activate’ its representational value. Hence the need for an ‘intellect’, i.e. a mental processing centre that coordinates all the relevant mental states and executes an orchestrated output. A powerful version of intellectualism is the traditional computationalist theory of mind, according to which mental states have a content (in the form of sentence-like representations) that the mind manipulates in accordance with certain syntax rules, and an attitude (i.e. a functional specification of how the sentence is to be processed)—the mind itself being a system with a memory storage (where all the mental states are stored) and a cognitive operator, or a central executive (which performs the computational operations on the mental states and issues in the behavioural outputs). Thus *mental states* are a representational content plus an attitude towards said content (e.g. belief, desire, intention), and *thought* (i.e. processes like deliberation, calculation, and planning) is the syntactic manipulation of such mental states. An *agent*, in this context, is most crucially identified with the central processing and executive system that manipulates the representational states by applying the syntactic rules, ultimately producing a behavioural output. And *action* is the behavioural product of syntactically processing representational mental states.³

Now, not every behavioural output counts as an action. In an intellectualist framework, the agent acts whenever the central cognizer’s manipulation of the right mental states properly produces a certain behaviour; behaviour does not count as an action whenever the manipulated mental states or the production process are not the right ones.⁴ There are rich debates in the intellectualist tradition concerning what are the right mental states (e.g. whether belief and desire suffice, or whether something like intention is also needed), and what the nature and role of each is. There also is great debate concerning what counts as properly producing a behaviour

---

³ For more detailed critical discussions of the varieties of representationalist theory, see Rowlands (2006); Gallagher (2008); Chemero (2009). Nowadays representational theories of action can be much more versatile than the traditional computationalist theory described in this paragraph, but we will get to that in a bit.

⁴ Davidson’s (1963) causal theory of action is a classical intellectualist account based on a computationalist model of the mind.
so that it counts as an action (on which see §2.2.3 below). But all intellectualist accounts of action share two fundamental properties, namely, that intentional actions are such because they are accounted for in terms of (1) the agent’s representational mental states, and (2) the mind’s centralized, cognitive manipulation of such states.

2.1.2. Anti-intellectualism

This core intellectualist view is not accepted by everyone, however. Over against the intellectualist framework, another perspective emphasizes intentionality’s directionality in opposition to its aboutness. From this alternative standpoint, a behaviour can count as an intentional action even if it was generated without any recourse to representational mental states and syntactic manipulation processes. Rather, this perspective identifies a behaviour as an intentional action if it reveals the agent as meaningfully directed towards its environment, independently of whether the meaningful directedness relies on mental representation or not. Action is not fundamentally mediated by mental content; in its primary sense, action is direct and immediate normative interaction with the world. The proponents of this framework often present their views as reactions to intellectualist accounts, which is why it makes sense to call it anti-intellectualist.

Now, what exactly constitutes an action’s directedness and normativity, if not the intellectual processes that gear behaviour toward an intended goal? What is the source of an action’s meaning, other than the contents of its associated mental states? Given that many of the current anti-intellectualist positions appeal to notions developed by Maurice Merleau-Ponty, it makes sense to spend a while reconstructing the main outlines of his views on intentional action.

The origins of anti-intellectualism

(A) Operative intentionality

Merleau-Ponty follows Husserl in calling “act intentionality” the kind of openness to the world that we have referred to as ‘aboutness’, which is most evident in activities like planning and deliberation. However, Merleau-Ponty considers act intentionality to presuppose a more basic kind of openness to

---

5 An intellectualist position with respect to action is expressed by what Bratman (1984; 1987, Chapter 8) calls “the Simple View”, according to which a behaviour \( \Phi \) counts as an intentional action if it is caused by an intention to \( \Phi \). Bratman criticizes this view, and leans toward the broader “Single Phenomenon View”, which states that a behaviour \( \Phi \) counts as an intentional action if it is caused by an intention, though not necessarily an intention to \( \Phi \). Both the Simple View and the Single Phenomenon view belong to the intellectualist perspective, according to which behaviour attains its intentional character because it is produced by the relevant cognitive manipulation of representational mental states like intention.
the world (what we have called ‘directedness’), which in turn makes the former possible. Husserl’s name for this non-representational intentional link with the world is “operative intentionality”. In order to explain the latter concept, it is useful to focus first on the intentionality of perception.

A given perceptual act or experience never shows more of an object than one of its profiles or faces. The object’s complete nature escapes the actual grasp of each individual perceptual experience, because the perceiver must always be situated in a particular standpoint, from which only some of the object’s aspects are accessible, many others remaining necessarily hidden. Though this is the case, perceptual consciousness nevertheless provides us with awareness of objects as unities, and not just with a bunch of disaggregated perspectives. How is this unity of perceptual objects possible? I.e. how does perceptual experience transcend the profiles to grasp the whole object in its multiplicity?

According to Husserl, the unity of the perceptual object is possible because of the temporal structure of consciousness. Each perceptual experience is not only a presentation of the object’s currently perceived profile, but also a retention of the profiles perceived in the immediate past, and an anticipation of the faces not-yet-perceived: the aspects of this object that have become visible to me allow for me to anticipate the other aspects that I am not seeing but may get to see in further perceptual experiences. This anticipation intrinsic to perceptual experience reveals the nature of operative intentionality. Even before I can have consciousness of an object as such (i.e. as some thing, as having a specific and specified nature that I can represent), before I have even collected the different perspectives of this object into a single unity, consciousness seeks to anticipate the faces of the world it has not yet perceived, and in this search for unity lies its originary directedness to the world. This is how, on the one hand, intentionality is at work prior to there being representations and representational manipulations, and on the other, such pre-representational directedness makes representational states possible: for without anticipation it would be impossible to form the most basic experiences of perceptual objects as unities.

---

6 About the idea of protention in perceptual experience, Husserl says:

For example, there belongs to every external perception its reference from the ‘genuinely perceived’ sides of the object of perception to the sides "also meant" not yet perceived, but only anticipated and, at first, with a non-intuitional emptiness (as the sides that are ‘coming’ now perceptually): a continuous protention, which, with each phase of the perception, has a new sense.

(1950/1960, §19, 82)

Here I take ‘anticipation’ and ‘protention’ to be synonyms. Notice that the case of perception is precisely that: just a case, an example of the phenomenon of anticipation, which belongs to the intentionality of all consciousness. For more on the relation between operative intentionality and protention, see Pachoud (1999).
Husserl and Merleau-Ponty call “operative intentionality” this originary directedness that aims at the unity of the object by going beyond the mere aspects given to consciousness.

Now, operative intentionality thus portrayed implies a certain normativity. The intentionality of consciousness is neither a passive receptivity of sensory data, nor a symbol-processing capability, but rather an active tendency toward grasping objects, in the sense of completing patterns: operative intentionality seeks the unity of the object via the satisfaction of its anticipations. Husserl referred to this purposive and originary motion toward the world as “a teleology of consciousness”.7 Under this light (and in a close analogy with Kant’s account of aesthetic experience), the operative intentionality of conscious experience is a sort of purposeless purposiveness, in the sense that consciousness aims at the world (hence it is purposive) without having a well-defined representation of what it is it aims at (hence its purposelessness).8 The aim of operative intentionality, despite being anticipated, is in fact necessarily indeterminate, because the object of anticipation is a potential future object. Hence, although through anticipation consciousness “intends the future in a more or less definite way, it is nonetheless open and indeterminate, for what it intends has yet to occur” (Thompson 2007, 360).

To go back to perceptual experience, when I perceive something I anticipate the different perspectives the current object might offer if I were to perceive it from different perspectives (further to the left or right, closer or further apart, from above or below, if I were to touch it, grab it, or stand on it…). That is, in perceptual experience I anticipate the faces that the object may offer given my spatiality and my possibilities of motion. Because perceptual anticipation is structured by my body’s potentialities for motion, operative intentionality, although prior to reflection and representation, nevertheless presupposes the spatiality and motility of embodied cognition. Along these lines, Merleau-Ponty develops an embodied and situated account of operative intentionality, a motor intentionality.

(B) Motor intentionality
For Merleau-Ponty, we actively intend the world in the first place by perceiving and moving through it. So sensorimotor activity is the originary locus of intentionality. The directedness of consciousness is thus revealed, and given concrete direction, in the way we move through space. With motor intentionality, anticipation acquires bodily dimensions. Anticipation becomes my

7 This teleology of consciousness reveals that the everyday uses of ‘intention’ and ‘intending’ are not entirely foreign to the phenomenological concept of intentionality (Pachoud 1999, 197). For a comparison between Husserl’s and Merleau-Ponty’s uses of this notion, see Dauenhauer (1979).

body’s tendency to be “geared onto the world”, which occurs “when my perception presents me with a spectacle as varied and as clearly articulated as possible, and when my motor intentions, as they unfold, receive the responses they expect from the world” (PP, 292).

Now we can describe operative intentionality’s directedness in normative terms, as the body’s tendency toward the “maximum sharpness of perception and action”. I intentionally move through the world searching for ways to confirm the perceptual unity I anticipate, and actualize it through my motions. The basic norm of motor intentionality is thus the tendency toward achieving what Merleau-Ponty calls maximum-grip, a concept that points to a maximization not only of my powers of motor action, but also of the environment’s capacities to solicit motor responses from my body (PP, 292). My body is here conceived not as the objective, material body, but as the lived body of sensorimotor experience, i.e. “as the potentiality for certain movements, as the demand for certain preferential planes” (292).

Some illustrations should help clarify the notion of maximum grip as the originary normativity at the basis of intentional action. (It makes sense to introduce them here too, because they are common points of reference for later anti-intellectualists.)

For each object, as for each picture in an art gallery, there is an optimum distance from which it requires to be seen, a direction viewed from which it presents most of itself: at a shorter or greater distance we have merely a perception blurred through excess or deficiency. We therefore tend towards the maximum of visibility, and seek a better focus as with a microscope. […] The distance from me to the object is not a size which increases or decreases, but a tension which fluctuates round a norm. (PP, 352)

Merleau-Ponty explains what he means by the painting “presenting most of itself” from the optimum distance in a later passage:

[T]here is a total logic of the painting or the spectacle, an experience of coherence between colors, spatial forms, and the sense of the object. A picture in an art gallery, when seen at an appropriate distance, has its internal lighting which confers upon each patch of colours not only its colour value, but also a certain representative value. Seen at too close quarters it falls under the prevailing lighting of the art gallery, and the colours “then no longer act in a representative manner, and no longer present us with the image of certain objects, but act as smears of paint on a canvas.”

A quote from Der Aufbau der Farbwelt by David Katz.
attitude and isolate part of the field, then the colour itself changes, and this green, which was meadow green, when taken out of its context, loses its thickness and its colour as well as its representative value. (PP, 364–5)

My body can relate to the meadow depicted in the painting as a meadow only from a certain distance, because seen from other distances the relationships between the painting’s elements vary so much that the ‘meadow-ness’ of the greens is lost, and transformed into just patches of green paint. Thus the unity of the painting as revealing a whole world to me is lost unless my body is geared toward the painting in the way in which it can reveal most of itself.

Merleau-Ponty’s interpretation of the painting example is worryingly representationalist, in the sense that he seems to think the only way for me to grasp the painting’s spectacle is by grasping the scene that the painting represents. However, this helps make his point by treating the painting as a world. Each object in my world, i.e. in my familiar practical environment, requires that I approach it from a certain distance, in a certain position, at a particular speed, and so on, in order for its unitary sense in the context of my practical life to be revealed and actualizable.

A quite different analogy from the artworld may perhaps further clarify the point. In creating Fountain, Duchamp did not do much more than rotate an everyday-use object and place it in an unaccustomed position in the middle of an unaccustomed situation. This gesture, however slight, forces the spectator to approach the object, spatially, in a way that cannot be the way in which it would habitually be approached. We can intellectually recognize what it is, but because we cannot grasp, in a motor way, its place in the current scene, it generates in us a tension that we cannot dissipate, because maximum grip on the scene is impossible.

This interpretation of Fountain in light of motor intentionality reveals that the operative normativity of perception works in the skilful manipulation of objects in our habitual spaces. In another illustration, Merleau-Ponty speaks about how a blind man’s stick is no longer an external object, but actually has become a part of the person’s body space, insofar as it extends his sense of touch and provides an access to the world analogous to sight (PP, 165–166).

Something similar occurs in the way a typist relates to his keyboard. What explains my skillfully typing a word, Merleau-Ponty holds, is not my representational intention of the word to be typed, which is computed to produce a representation of the correct movements for typing each word, which are then translated into motor responses, but rather my “motor intentions” (i.e. my
search for maximum sensorimotor grip) realized by means of my body, which has been extended through habituation and now includes the keyboard itself (PP, 167).

It is revealing that Merleau-Ponty makes no distinction between the words ‘habit’ and ‘skill’. There is at least a historical reason for this procedure (the words ‘habit’ and ‘ability’, just like the French ‘habitude’ and ‘habilité’, have a common etymological root), but Merleau-Ponty seems to do this for a philosophical reason (Dreyfus 1996): by keeping habit and skill together he emphasizes the idea that the lived body is not just an object in space, but rather inhabits space through its familiar sensorimotor activities. Inhabiting space implies both the habituality of repeated movements in familiar locations, and the skill of satisfactorily negotiating those habitual spaces through motor intentionality.

Merleau-Ponty more forcefully makes the same point (i.e. that my body intends the world through its habitual/skilful sensorimotor activity) while discussing the case of an organist who could play his repertoire in an organ that has a set of manuals and stops different from the ones he regularly uses after only an hour-long practice session (PP, 167–168). The organist’s ability to play his habitual set in an unfamiliar instrument illustrates how clarifies the nature of habit itself. This case, in fact, shows that habit is neither a blind linkage between sensory input and behavioural output (if this were the case, a one-hour practice session would not suffice to override old input-output correlations and develop new ones), nor the application of a plan reflectively developed by processing newly-acquired representations of the instrument’s features, and their objective spatial relations to one another (this mediation of the representations and reflections between the music and the agent’s motions would be impossibly complex, and yield an unnatural performance). Rather, Merleau-Ponty holds, habit is the opening up of a motor space, which the embodied agent can inhabit by skilfully moving through it so as to achieve the “maximum sharpness of perception and action”. Motor habit/skill is thus revealed as “an extension of existence”, and perceptual habit/skill, as “the coming into possession of a world” (PP 176–7).

The point can be made that habit/skill and motor intentionality are quite closely related, so much so that all motor intentionality may rest on a skill, and all skilful performances may be best described as non-intellectual intentional actions. This is in fact the view that recent proponents of anti-representationalism have put forward, and I will proceed to sketch it briefly.

---

10 Studies suggesting that skilled typists have poor explicit knowledge of the relative position of the individual keys in the QWERTY keyboard (Liu et al. 2010) corroborates the claim that typists do not use representations of the keyboard while typing. If representations and calculations were used, one would expect skilful performance to be correlated with the relevant explicit knowledge. For more on typing and skilled control, see §3.3.2 below.
Before this, though, I would like to provisionally confront an especially salient objection to the possibility of non-representational intentional action.

(C) Anticipation vs. representation

Merleau-Ponty argues against views that remove all intelligence from habit, or that make habit fit into an intellectualist account of intelligence, and holds, rather, that

the phenomenon of habit […] prompts us to revise our notion of ‘understand’ and our notion of the body. To understand is to experience the harmony between what we aim at and what is given, between the intention and the performance (PP 167).

This passage should be understood under the light of the notions of ‘anticipation’ and ‘operative intentionality’ discussed above. In a fully non-representationalist way, ‘intention’ and ‘understanding’ involve not the subsumption of a manifold of experience into a representational unit, or a symbol that semantically tracks an object, but rather the harmony between the anticipated aspects of the world and the sensory experience that we acquire through our motion, i.e. the harmony between the motor intention and the body’s sensorimotor activity.

Thus, the whole nonrepresentational account of intentionality is based on the idea that the anticipation included in each mental act is not a representation. But why not? After all, anticipation seems to be a sort of image or symbol of the object of experience, since it does not entirely fit the precise content of the present experience, but rather subsumes the present and past faces of the object it into a ‘predictive model’ of the object… why not consider this predictive model a representation? Isn’t the distinction between anticipation and representation somewhat artificial? And if so, wouldn’t the distinction between operative and act intentionality be artificial as well?

A reply to this begins11 by pointing out that, from Merleau-Ponty’s perspective, consciousness is not a collection of discrete time-slices (like each one of the picture-frames in a movie reel), but rather a continuous dynamic process, in which each moment of the process unfolds from the previous one and fades into the next. This is possible only by the temporal structure of what Husserl called the “living present”: each intentional experience is constituted not only by the presently intended object, but also by the “retention” of immediately previous objects and the anticipatory “protention” of subsequent objects. But neither retention nor

11 For the full story, see Gallagher’s (2008) discussion of recent minimal-representationalist proposals.
protention qualify as representations. Showing this requires introducing some of the features that representations traditionally have, and that neither retention nor protention do.

In a minimal description, for something to count as a mental representation it must (1) be mainly internal to the agent (probably instantiated as a neurophysiological organization inside the brain), (2) bear content that is external to itself, and (3) be decoupable from its original presentational situation. However, motor anticipations have none of these traits. For one, (1) they are not within the agent, because they are part of the ongoing motor coupling between agent and world. Of course, they must have a neural component, but the agent’s intentional reach towards the world does not stop inside the agent (let alone the agent’s brain), as the examples of the gallery-goer’s movement, the blind man’s stick, and the typist’s keyboard make clear. Motor anticipation is the process whereby my body seeks maximum grip through perception and motion, and therefore is not mainly located in one of my internal states, but rather in the coupling dynamics between agent and environment. Further, (2) motor anticipation does not offer a content that mediates between agent and world; rather, as just mentioned, it is a direct motor engagement with the world—it is not about the world, but rather in the world. And finally, (3) it would be hard to see how the anticipations that allow the typist to skilfully reach towards the keyboard, or the organist to play the organ, can be decoupled from the agent’s current position in front of their instrument, the particular instrument itself (since many of these anticipations are not transposable to slightly different instruments), and the agent’s concrete practical situation; but symbols must be detachable from their originary environment in order to play the role of mediators. Thus anticipation does not seem to be the kind of thing that could count as a representation.12

So far, I have characterized an intellectualist framework (which emphasizes intentionality as aboutness and considers intentional action as behaviour that is accounted for in terms of intentional states and the cognitive processing of such states), and contrasted it with Merleau-Ponty’s anti-intellectualism (according to which the originary kind of intentionality is motor directedness, and intentional action is intellectually-unmediated sensorimotor coupling with the

---

12 Two things are worth pointing out. First, motor anticipations can be interpreted as involving representations, if ‘representation’ is taken to mean an emulation of the body-environment interaction used to provide expectations of sensory feedback—a view that is widespread in motor control research (cf. Grush 2004). However, such body-environment emulations do not satisfy the decouplability condition, and would thus not count as canonical representations. Second, admittedly, a motor anticipation may be transformed into a representation (say, in the form of a rule about how to play the organ, or about where to stand with respect to a painting in relation to its size), but once this happens, the anticipation no longer plays a role in directing motor action, since now it would require a reflective process of planning or deliberation in order to be operationalized.
world that tends toward maximum grip). This is the landscape of the views on intentional action in its broadest terms (see Fig. 2.2).

Motor intentionality has been present in recent cognitive-scientific discussion partly through Hubert Dreyfus’ phenomenological account of skilful action; the enactive approach, which combines motor intentionality with insights from dynamic systems theory; and the extended mind theory, that explains cognition as inseparable from the agent’s active environmental coupling. I will briefly sketch these positions’ main outlines.

Contemporary anti-intellectualist perspectives

(D) Skilful coping as intentional action

Hubert Dreyfus has criticised traditional intellectualism, defended the relevance of non-intellectual intentionality for understanding human agency, and thereby become the main target of many new, refined intellectualist conceptions of intentional action (see §2.3 below). Dreyfus has extensively argued for the view that motor intentionality must be invoked in accounting for what he calls “absorbed skilful coping,” i.e. habitual and skilful bodily actions. As we acquire a skill, Dreyfus contends, we rely less and less on explicit rule-based cognitive processing, and more on affective responses and intuitions, through which we can perform discriminations of the situation’s relevant features so fine-grained that they would be impossible to achieve via rule-based cognition. Thus the more experienced we become, the less we reflect, and the more the situation itself appears to us as requiring specific responses from us. So motor intentionality can explain skilful coping, and reflective intentionality cannot, because only the nonrepresentational, emotionally-loaded responses can give us the level of fast and detailed and efficient discrimination required for expert action.

A crucial aspect of Dreyfus’ position is that expert, skilful agents do not really have to do much in order to discern what must be done and how to do it. In this sense, expert action is very
similar to a mere reaction. But this reactivity is not mere passivity, like that of a compulsion. Rather, it is the passivity of “letting myself be moved by the gestalt tensions I experience” (Dreyfus 2002a, 380). The latter passivity has an agential component, because although the movement does not originate in the agent’s occurrent mental states, it is possible only through the agent’s skilful sensorimotor coupling with the world. Such coupling is based on the expert’s prior experiences, which shape the way the situation looks and feels to her, thus soliciting a determined, immediate response. Dreyfus holds that such unity of past and future experience, of moving body and environment, is what Merleau-Ponty’s habit/skill aims at.

Against intellectualism’s core claims, Dreyfus holds that skilled action is not about constructing an inner mental representation of the world and then computing how to act on it. This would be very inefficient when the time is short and the situation is complex. Instead, we rely on our habits/skills to signal an immediate response to the world; so no inner model and no central executive are needed.

(E) Sensorimotor agency in the enactive approach
The enactive approach to cognition and action radicalizes Dreyfus’ approach to skilled action by generalizing a model of skill-based operative intentionality to all of cognition and all of action. They build the account from the ground up, using as a foundation a precise conception of biological autonomy, originally proposed by Francisco Varela (1979). Enactivism holds that “intentionality arises from the operational closure of an autonomous system” (Thompson 2004, 388). A system is autonomous if it generates and maintains its existence through its own activity. Unicellular organisms provide a paradigmatic example: they build and maintain the boundary that separates inside from outside (self from environment), and the maintenance of that separation also allows them to carry on their constitutive, internal metabolic processes. An autonomous system’s operational closure is the production and regeneration of the boundary and the inner processes that keep itself and the boundary in existence.

Why should intentionality arise from operational closure? First, because by distinguishing inside from outside through its own activities, the system enacts its own environment: it specifies its realm of interactions. And moreover, all autonomous systems need to stay in a continuous flow of exchange with their environments, but in a way that is adaptive, i.e. a way that avoids threats to their self-maintenance and maintains the autonomous organization under precarious conditions (Di Paolo 2005). A cell must manage the inflow and outflow of materials through its semi-permeable boundary; an animal must be on the lookout for the sources of nutrition that keep its metabolic processes going. These are cases of
autonomous systems retaining their adaptivity through environmental interactions. The need to maintain adaptivity is a sort of biological imperative that colours the organism’s world with *valence*, a valence that “is dual at its basis: attraction or rejection, approach or escape” (Weber & Varela 2002, 117). When the system’s interactions with its environment move away from its viability conditions, they are endowed by the organism with relevance, and appear as salient tensions and deviations from a norm; and movements away from the limits of viability are sensed by the organism as a relief of such tension.\(^\text{13}\)

For enactivism, therefore, openness to the world is grounded in the organism’s need to guide its environmental coupling toward situations of viability. This originary cognitive relationship receives the name ‘*sense-making*’ (Di Paolo 2005; Barandiaran et al. 2009). Sense-making itself, understood as the active sensorimotor guidance of environmental interactions (away from the limits of viability, and toward situations of stable autonomy-maintenance) is also the primary kind of intentional action, since in order for an autonomous agent to avert threats it must skillfully navigate its environment. Cognition and sensorimotor action are thus intrinsically linked, cognition being essentially “viable sensorimotor conduct” (Thompson 2004, 387).\(^\text{14}\) Enactivism thus echoes Dreyfus’ concept of skilful coping by holding that “our primary way of relating to things is neither purely sensory and reflexive, nor cognitive or intellectual, but rather bodily and skillful” (Thompson 2007, 247). But here skilful coping is radicalized in the form of sense-making to account not only for expert skill, but for the originary forms of cognition and action.

Operative intentionality, understood as sense-making, is the originary kind of openness to the world, and other forms of intentionality are derivative upon it. It differs from the act intentionality proposed by intellectualism because it does account for meaning and experience on the basis of representations (it rather accounts for both meaning and representations by reference to the autonomous system’s enacting self and environment through its interactions with the world), and it does not appeal to a central executive to coordinate sensorimotor action (enactivism appeals to dynamic systems theory to model how intelligent cognitive behaviour emerges as a result of lower-level processes). A challenge for the enactive approach is to account for higher-order behaviours of increasing cognitive complexity (like argumentation and planning) without the

---

\(^\text{13}\) This analysis has much in common with Aristotle’s conception of the locomotive soul discussed above (§1.1.2).

\(^\text{14}\) Importantly, the relationship between autonomous system and world is intentionally directed and meaningful, not as defined from an observer’s third-person perspective, but rather as determined by the organic system’s own norms of interaction with its environment, its own needs and its risks, as defined by its internal constitution (Barandiaran et al. 2009). Thus the notion of sensorimotor agency has much in common with Aristotle’s view of the locomotive soul discussed earlier (§1.1.2).
recourse to inner representations or a central executive controller. Work in the frontiers of enactive research addresses exactly that issue, working its way up from the cell to society.\(^{15}\)

\textit{(F) The extended mind}

The theoretical framework known as ‘the extended mind’ develops some of Merleau-Ponty’s insights about habit/skill in an interestingly different direction.\(^{16}\) Recall how Merleau-Ponty describes the blind man’s stick as no longer being an external object, but having become a part of his “body space” that provides an access to the world analogous to sight; and how he treats parts of the environment that contribute to habitual dynamics of action and cognition (like the organist’s keyboard) as “an extension of existence”. The extended-mind theorist unpacks these ideas by suggesting that some cognitive processes do not take place entirely inside the body or the skull, but transcend this boundary to include elements of the agent’s environment that actively contribute to them. Without their contributions, our cognitive and agentive capacities would be radically reduced.

Mathematical operations provide a classical example. We are quite good at quickly performing simple operations (like $5 \times 5$), but rather terrible at performing more complex ones (like $5555 \times 5555$). To solve problems like the latter, most people would rely on pen and paper, or a calculator, or an abacus, or some external medium. The external device becomes a temporary memory storage, where we deposit the solutions to each one of the simple sub-operations, and this enables us to construct the final answer. The fact that without the pen and paper most of us would struggle to solve the problem, suggests that our interactions with them play an active part in the cognitive process of multiplication. (And even those who do not struggle seem to succeed by having internalized the pen-and-paper structure, and deploy it imaginatively in their train of thought, which also shows the augmenting power these tools have.\(^{17}\))

The notion of scaffolding embraces a wide class of physical, cognitive, linguistic, and social structures that augment our cognitive-practical capacities, allowing us to achieve goals that we would otherwise find unachievable (Clark 1997, 194–195). To see the relevance of scaffolding, think about how remarkably structured an average day of your life is: how neatly connected it is with your long-term plans. Now imagine how a day in your life would be without

\(^{15}\) The enactivist gradually expands her account from the realm of biological sense-making that is purely determined by the need to retain adaptivity, to more complex kinds of intentional behaviour where biological needs no longer exhaust the normativity in play, like affect, intersubjectivity, and social interaction (see De Jaegher & Di Paolo 2007; Di Paolo 2009; Froese & Di Paolo 2011).

\(^{16}\) See Clark (1997); Clark & Chalmers (1998).

\(^{17}\) Heath & Anderson (2010, 236) tell about older Chinese merchants who move their fingers as though moving the beads of an abacus to perform complex calculations. This is another case of internalized scaffolding.
calendars, computers and emails, telephones, offices, public transit, bureaucracies and your specific role within them, colleagues, social norms regulating public space, and so on. Even more radically, imagine a day without language and the ability it gives you to coordinate behaviour with others, and within yourself through inner speech. How much of that long-term structure could be achieved without the environmental scaffolding around you? A likely answer is: nothing, or close to nothing.

So external scaffolding seems necessary for achieving many of our goals, *a fortiori* when we consider the architecture of the mind. Evidence suggests the human brain is less like a serial computer (with a memory that systematically stores internal states and a central processing unit that can access, process, and thread them together), and more like a neural network (i.e. a large network of interconnected nodes whose patterns of co-activation are shaped by experience).\(^\text{18}\) This explains why brains have a very limited capacity for serial processes that coordinate multiple bits of information coming from different sources (say, perception, memory, volition, etc.) into a coherent plan that is translated into motor routines. Rather, we seem to work on the basis of automatic, semi-independent parallel processes (shaped phylogenetically by evolutionary pressures and ontogenetically by acquired habits), and to be very bad at integrating these processes into complex, coherent patterns, in part because multiple parallel processes can interfere with one another. Brains are good at quickly completing simple patterns, but not too great at structuring complex, stepwise chains of thought.\(^\text{19}\)

If this is so, then the big puzzle for neural-network-based cognitive systems (us included) “is how to maintain coherent behavior patterns as the systems grow more and more complex and are required to exhibit a wider and wider variety of behaviors” (Clark 1997, 33). Scaffolding is crucial in achieving this coordination, because it not only provides memory storage (as in the pen-and-paper case); it also allows us to *reframe information* so that relevant patterns become identifiable (like when a Tetris player turns the piece around in the screen to determine where it fits best, or a writer reorganizes her notes while writing an essay), to *minimize cognitive effort* in guiding action (like when we leave the vegetables visible and close to the fridge door, while hiding the snacks deep in the counter; or like standard signalling in public

---

\(^{18}\) See below (§2.2) for more on associative neural networks, and the differences between them and the traditional computational models of the mind.

\(^{19}\) “Good at Frisbee, bad at logic” (Clark 1997, 60) seems to sum it up.
spaces works for cognitively efficient guidance); or to coordinate action within ourselves and with others (especially through the use of language).

The extended-mind theorist argues on these grounds that much of the cognitive resources required for complex, intelligent, long-term action are not available in the brain, but are outsourced to elements in the environment. She thus rejects the intellectualist view that intentional action requires an internal executive controller that centralizes and structures multiple threads of cognition and behaviour. Executive control is a product of environmental scaffolding. The crucial insight is that a behaviour’s complexity need not imply the involvement of a centralized executive planner operating on the basis of a detailed representation of the situation and coordinating each individual process toward a goal. Complex, goal-oriented action may also emerge as a result of distributed processing, in which the agent performs several associative pattern-completing routines, and these are both constrained and augmented by the networks of external scaffolding that includes tools, institutions, social norms, and linguistic interactions. If this is so, then there is no need to posit a central executive processing subsystem.

Just because humans can do logic and science, we should not assume that the brain contains a full-blown logic engine [...]. Instead, both logic and science rely heavily on the use and manipulation of external media, especially the formalisms of language and logic and the capacities of storage, transmission, and refinement provided by cultural institutions and by the use of spoken and written text. [...]. The brain need not waste its time replicating such capacities. Rather, it must learn to interface with the external media in ways that maximally exploit their peculiar virtues. [...] Gone is the central executive in the brain—the real boss who organizes and integrates the activities of multiple special-purpose subsystems. [...] In place of this comforting image we confront a vision of mind as a grab bag of inner agencies whose computational roles are often best described by including aspects of the local environment (Clark 1997, 221–222).

Despite rejecting the core central-executive tenet of intellectualism, extended-mind theorists accept (a modified version of) the representationalist claim. The classical-intellectualist view that

---

20 Clark considers public language to be the ultimate scaffold, whose function far exceeds mere communication. This is, on the one hand, because its nature as a code that is context-independent (word meanings tend to stay the same through different situations), modality-neutral (neither visual nor auditive, but both), and easy-to-learn, allows for us to offload a great amount of memory content onto the structure and content of language. And on the other hand, because it opens the door for second-order, recursive thinking (e.g. the passage from ‘the cat is on the mat’ to ‘She believes the cat is on the mat’ to ‘I think she believes the cat is on the mat’ and so on), which can be seen as the foundation for mathematics and for the purposeful design of scaffolded environments. For more on this, see Chapter 5 below, Clark (1998) and Heath (2014, 56).
we have a detailed inner representation of the world, that we manipulate in order to act, is not ecologically viable, since the computation processes required for building, updating, and translating this representation would be too taxing and slow to yield viable results—the lion would eat us before we could find shelter in the cave (Clark 1997, 21–22). However, internal representations are still needed in order to perform “representation-hungry” operations, which require our tracking a complex, multi-modal object that is not directly available (e.g. complex imagination, counterfactual reasoning, and planning). However, the required representations are much more partial and specific than traditional intellectualism would have it. Classical representations are action-neutral reflections of the world; extended-mind representations are “action-oriented”: partial models that track specific action-relevant situational features, and that are updated or discarded as the agent’s practical needs evolve. Because these representations are already geared toward action, they require much less cognitive effort to be deployed. External scaffolding greatly reduces the need for representations because exploiting external or public symbol structures (like language) implies that most of the representations needed for problem-solving are already built into the resources used to frame the problem itself (Clark 1997, 200). That said, some higher-order processes must necessarily deal with (action-oriented) representations.

**Uniting the anti-intellectualist front**

There are, therefore, two main fronts of anti-intellectualist opposition to intellectualism. One of them rejects both the representational and the central-controller tenets of intellectualism. This stance is shared by enactivism (which considers that the notion of representation is unable to account for cognition, because it fails to explain the origins of meaning and relevance) and the views sometimes called *radically embodied cognitive science* (which rejects representations, and therefore central-controller theories, altogether), both of which employ dynamic-systems tools to model the mind. This stance is also shared by Dreyfus’ skilful-coping theory, which uses neural-networks models but also rejects both intellectualist tenets.

The other anti-intellectualist front is mainly represented by the extended-mind approach, and is usually considered as more of a central position because, despite its rejection of the central-executive claim, it accepts the need for inner representations. To many in the anti-intellectualist ranks the extended-mind approach is not radical enough, and they consider the

---

21 The idea that objective representations are too slow to be ecologically viable is derived in part from attempts to construct artificial intelligence in accordance with that model. Clark describes some such attempts that did not achieve viability, despite being endowed with remarkably large amounts of subject-independent information about the world.
23 More on minimal (as opposed to classical) representations below (§2.2.1).
only way to go forward is rejecting all things representational. From the extended-mind side, however, there have also been criticisms of the plausibility of radical anti-intellectualism.

Despite the valid disputes about the need to incorporate, reject, or reconceptualize representations, in this work I will focus on the larger debate, i.e. the question about whether the production of intentional actions requires the involvement of an inner, central executive; or whether (at least some) intentional actions can be produced without the involvement of such coordinating capacity, and the lower-level, automatic, pattern-completing mechanisms are sufficiently intentional to produce intentional action on their own. To defend anti-intellectualism at this level, I submit that both sides of the anti-intellectualist front have a better case to make if they stand together in facing the powerful, novel intellectualist positions that I am about to introduce. Like Clark, I think this is a discussion in which progress demands cooperation rather than competition. The open intra-group debates (about whether to redefine or reject representations; about how to properly model the mind’s architecture; about whether the internalism–externalism language is useful at all; about the differences between extending the mind and incorporating something within the cognitive system) can be developed in parallel with the inter-group dispute against intellectualism, which is a pressing concern and a common interest of both the radical and moderate factions. For it is not clear that the very first step away from intellectualism, the rejection of a central executive, is granted at all.

---

24 See e.g. Thompson & Stapleton (2009) and Chemero (2009).

25 Rowlands (2009) argues that enactivism is incompatible with the extended mind perspective, and that the former may be an altogether incoherent approach.

26 “Such radicalism […] is both unwarranted and somewhat counterproductive. It invites competition where progress demands cooperation. In most cases, at least, the emerging emphasis on the roles of body and world can be seen as complementary to the search for computational and representational understandings.” (Clark 1997, 149)
§2.2. Intellectualism and anti-intellectualism about associative actions

The Terminator T-800 is quite a powerful machine. Equipped with advanced AI systems and a strong, titanium-alloyed endoskeleton, it is an incredibly efficient destruction device. But the newer, T-1000 model seems tougher in every way. At least equally as smart, but made of liquid metal, it can restructure itself to take the shape and appearance of anything it touches, even imitating people’s facial expressions (something rather impossible for the T-800). This makes the T-1000 more able to stand for a flesh-and-blood, organic human being.

The anti-intellectualist ideas have presented great challenges to the classical, intellectualist view of the mind. In fact, it could be said that the relevance of embodiment, situatedness, and embeddedness for human cognition, together with the (now more evident than ever) information-processing limitations of the human mind, make the traditional, cognitivist framework, with its centrally controlled, rule-based computations, all but obsolete.

Or do they?

In many ways, modern intellectualism is more like the shapeshifting Terminator T-1000 than like the old, rigid T-800. As I attempt to show in this section, intellectualism can incorporate into its own framework many of the anti-intellectualist insights about embodiedness and situatedness, to the point that the new, situated variety of intellectualism seems (prima facie at least) to have at least the same explanatory power as its the anti-intellectualist opponents. As mentioned several times (§2.1.2), one of the main objections anti-intellectualism raises against intellectualism is its inability to accommodate fast, fine-grained, quick-and-dirty, skilled, expert, and habitual action and cognition. But just like the newer T-1000 can replicate any of its

---

opponents’ behaviours, there may be nothing that the newer situated intellectualist model cannot do, no matter how embodied, situated, and embedded.

The core of the dispute is precisely those actions based on associations rather than rules, like playing Frisbee rather than doing Logic. For the crucial point of opposition is that anti-intellectualists think it is unnecessary for such actions to be coordinated by a central controller, but situated intellectualists consider all intentional actions to have the central controller’s activity as part of its production. So which side can do more justice to the evidence we have concerning associative, skilled, and habitual actions?

In this section, (2.2.1) I introduce the concept of situated intellectualism, a kind of intellectualism that adopts situated and embodied insights by incorporating neural-networks models and the concept of associations, i.e. mental states with non-propositional, associative content. After this is done, (2.2.2) we can review the whole landscape of intentional action described thus far, and the main debate. Finally, (2.2.3) I set the stage for the following investigation by reviewing how situated intellectualists account for the intentionality of associative behaviour by attempting to specify the control that top-down executive processes can exert over it, whereas anti-intellectualists propose a reflection-independent account of the automatic intentionality.

### 2.2.1. Situated intellectualism

**Associative mental states**

Traditional representationalism holds that representations are inert contents, stored in the form of conceptual and propositional structures, that the mind must actively interpret by means of reflection in order to successfully engage with the world. From this perspective, representations are more or less agent-independent descriptions of the world, whose relevance to our goals and intentions we must cognitively figure out in order to produce coherent and well-adapted action. Representations *describe*, and then cognition *prescribes* (by means of reflective processes like counter-factual reasoning, deliberation, and planning). Finally, intentional action is the behavioural execution of cognition’s prescriptions.

With the help of anti-intellectualism, it is easy to see that this division of labour between representation and cognition will fail to provide a sufficiently robust account of action. It puts too much reflective load on the agent for the whole system to be viable in highly demanding

---

28 For now, I use the word ‘reflection’ to refer to any cognitive process that requires the participation of a central executive controller. I will refine the conception of reflection in the next Chapter (§3.2).
contexts that require fast and adaptive responses—in short, propositional intellectualism seems unable to account for skilful coping. Thus Dreyfus defended his anti-representationalist intellectualism by arguing that skilful coping was not explainable in terms of propositional representations (2002b, 419).

But from the rejection of traditional, propositional intellectualism it does not follow that we must endorse an anti-representationalist account. For the representationalist can let go of an exclusively propositional view of representations without letting go of representations altogether. And once she lets go of such restrictions, a whole new world of possible representations opens up. Representations need not be prescriptively mute: they themselves can be both descriptive and prescriptive, both informing the agent about the world and providing her with guidance about how to act. Representations, in a word, can be conceived of as associations between descriptive and prescriptive content. This generates a new form of representationalism, much closer in spirit to anti-intellectualism (since it minimizes the need for reflection), but still holding a fully representational view of intentionality and the existence of a less busy central controller. I call this position situated intellectualism.

Tamar Gendler’s account of ‘alief’ is a recent intellectualist theory of just this kind.29 Gendler starts out by reporting cases in which subjects experience cognitive conflict. One of them is a study by Paul Rozin, in which participants were given a square of high-quality chocolate fudge, and then presented with two more pieces of the same, one shaped like a muffin, and another shaped like “a surprisingly realistic piece of dog feces” (Rozin et al. 1986, 705). Although they knew that both pieces were made of the same fudge they tried before, participants reported a significantly higher preference for eating the muffin-shaped piece rather than the poop-shaped one. There is a cognitive tension here between the belief that both pieces are made of the same chocolate fudge and the aversion towards only one of them. Gendler reports several analogous cases of cognitive tensions, and argues that the best way to account for them is by positing a non-propositional kind of representational state that can conflict with traditional, propositional mental states. She labels this kind of state ‘alief’, to highlight its opposition to belief, and to emphasize that these mental states are automatic (i.e. immediately activated by the presence of an environmental stimulus independently of the agent’s reflective

---

29 See Gendler (2008a; 2008b). Gendler’s ‘alief’ is not the first philosophical conceptualization of non-propositional representational states that are at once descriptive and prescriptive; Cussins’ (1992) “cognitive trails”, Millikan’s (1995) “pushmi-pullyu” representations, and Clark’s (1997, Chapter 8) “action-oriented” representations are particularly relevant precursors. I focus here on Gendler’s alief because of its direct links with the psychological literature on automaticity that will be the topic of the next Chapters, and because, although adopting an associative account of representations, it remains intellectualist (as opposed to other thinkers who endorse associative representations, like Clark).
states); associative (i.e. clusters of descriptive, affective, and behavioural representations that have become linked to one another through habituation); and arational (i.e. recalcitrant and impervious to new evidence and reflections).

The associative character of these mental states is worth reflecting on. In contrast with traditional, propositional mental states, which, as mentioned above, are usually understood in terms of the computer model of the mind (i.e. as internal symbolic states that are read by a central processor and processed in accordance with a set of syntactic rules), associative mental states are modelled after neural networks. These networks are composed of large numbers of nodes (analogous to neurons) which have connections of varying strengths or weights to other nodes (analogous to the synapses between neurons). The weights get strengthened or weakened by the network’s accumulated activity (analogous to learning). Through experience, associations between diverse nodes emerge and strengthen, or weaken and disappear. When one node is activated, this activates the nodes that are positively associated with it, thus forming a pattern of activation that has been shaped through the system’s past neural activity. The situated intellectualist interprets such patterns of activations as representational mental states.

Why are associations relevantly different from propositional mental states? Four things are worth mentioning. First, associations do not require rule-based computational processing: all the processing required is built into the structure of activation patterns, which implies that, insofar as behaviour relies on associations, computational effort is off-loaded to past experience and the practical environment. After having had negative experiences with knives, for example, the nodes in my nervous system that represent knives become associated with those that activate the feeling of danger, and with the nodes that activate the caution motor response. Thus, whenever I perceptually encounter a knife I now automatically feel there is danger and have a tendency to move more cautiously than usual. In this case, producing situation-tailored behaviour does not require centrally-controlled cognitive processing.

The knife example fits nicely with Gendler’s claim that aliefs have an R-A-B structure: paradigmatically, aliefs are associations between a Representational component, an Affective response, and a Behavioural output. This is the second aspect of associations worth highlighting: paradigmatic associative representations bind descriptive, affective, and prescriptive components together. From this perspective, perception does not gather neutral data that must then be interpreted, and then acted upon; rather, the relative salience of different perceptual features of a

---

30 This is a rough sketch of how ‘connectionist’ systems work. These activation patterns may or may not be interpreted as representations. More on this below. For more on connectionism, see the papers collected in Clark & Lutz (1992).
scene is partly determined by the agent’s past experience, and its apprehension immediately activates associated affects and prescriptions about how to behaviourally respond.

Third, **associative states are not immediately modifiable by the acquisition of new information**. Beliefs can be updated remarkably quickly with the acquisition of new information, in a way that aliefs (and associations generally) cannot: if I find out that my favourite neighbourhood pizzeria has moved to a new location, I immediately update my belief about its location to reflect this (so that if somebody asks me “Where is the pizzeria?”, or “Where can I find the best pizza around?” I can offer a correctly updated reply); but even so, the next time I crave for a pizza I may end up walking to the old location out of habit, i.e. because of the over-practiced association between the pizza’s representation, the craving, and the motor behaviour that leads to the habitual place. Propositional mental states are updated and modified via rule-based, syntactic processes; but associative mental states are shaped through repetition and habituation. So in order to rectify my association between the pizza and the old place, I will have to keep a close, cognitively effortful watch over my behaviour the next few times I head over to the pizzeria, until a new habit emerges and substitutes new associations for the old ones. In Gendler’s vocabulary, associations, unlike propositional mental states like beliefs, are “evidence-insensitive” (2008b).

Finally, **associative networks can process semantic information, but not syntax**. They activate representational elements that have become semantically related (e.g. ‘Dog poop! Gross! Stay away!’, or ‘Knife! Sharp! Pain! Watch out!’), but they cannot process these semantic relations in accordance with the rules of logic (cf. Levy 2014). In another one of Rozin’s experiments, also cited by Gendler, participants saw how the experimenter filled two empty bottles with sugar from the same source; then they were asked to attach to one bottle the label *Sucrose, table sugar*, and to the other the label *Not sodium cyanide, not poison* “with a red skull and crossbones preceded by the word *Not*” (Rozin et al. 1990, 383). The experimenter proceeded to sweeten one glass of Kool Aid with sugar from the *Sucrose*-labelled sugar bottle, and another glass of the same Kool Aid with sugar from the *Not-poison*-labelled sugar bottle. As in the chocolate fudge case, although the participants knew that the sugars and the drinks came from the same sources, they were nevertheless reluctant to drink from the glass that was sweetened with the *Not-poison* sugar. Gendler infers from this that participants experienced the same kind of conflict, this time between the belief that there is no poison in the drink, and the alief ‘Poison! Cyanide! Dangerous! Avoid!’ This reveals that associative networks are capable of processing semantics
(like ‘poison’, ‘cyanide’, and the red ☠️ symbol), but are insensitive to syntax (like the logical negation).\footnote{That is Levy's (2014, 34) interpretation. Gendler expresses this point by saying that aliefs work in “hyperopaque contexts”, i.e. contexts that do not permit \textit{salva veritate} substitution of two expressions recognized by the subject as co-referential: “Even if I believe that the phrases “not poison” and “safe to consume” pick out coextensive classes of substances […], still the aliefs activated by the two expressions may be wildly dissimilar" (Gendler 2008a, 650).}

In short, associative networks are great for pattern recognition tasks, like motor control, face and handwriting recognition, and the like—tasks that we find ourselves to be very fluent at. But they are far from great (indeed mostly incapable) at tasks that require sequential processing, like logical reasoning. And sometimes the need for sequential processing may be hindered by the activation of the automatic pattern-recognition routines.

\textit{Situated intellectualism}

The introduction of neural networks and associations was received by some as revolutionary, because they were sometimes interpreted as a move away from intellectualism. Offering a situated, action-centred, reflection-free model of cognition and action, they were thought to lead to the rejection of both the central-executive and the representationalist tenets of intellectualism. But they did not.

Dreyfus seems to have thought that the association-based model led to anti-representationalism, since there was no need for propositional representations to account for habitual action and skilful coping.\footnote{He says, e.g., that “skills are, at some level of explanation, nonrepresentational. But […] I don’t mean non-intentional; I mean non-propositional” (Dreyfus 2002b, 415).} But it did not, because associations can be seen as just another type of representations. Dreyfus claims that a particular neural-network model of the mind, Freeman’s (1991) attractor theory, is isomorphic with Merleau-Ponty’s non-representational account of motor intentionality; but he recognizes that Freeman himself expresses his theory in representational terms (Dreyfus 2002b, 419–20). He does not believe that activation patterns (“chaotic attractors” in Freeman’s terminology) should be interpreted as representations, although he admits they can. And Freeman is not alone in this; in fact, most neural-networks researchers have a representationalist interpretation of associations (e.g. Smolensky 1988).

And Clark thought that modelling the brain after neural networks would lead to the demise of intracranial central executive centres: the brain’s parallel, associative processes would acquire coherent temporal structure thanks only to external, environmental scaffolding (see §2.1.2F above). But it did not. One crucial reason for this is the very kind of cognitive conflict depicted by Gendler: standing on the glass floor of Toronto’s CN Tower makes many of us
shiver in fear, despite the fact that we know we are safe, so much so that we can override the fear
and stand on the invisible floor. The ability to override associations on the basis of beliefs and
new evidence does not seem to be a task that neural networks can perform. It seems to require a
sort of ‘central controller’, of arguably limited capacity, but still capable enough to resist
temptations, overcome fears, break habits, and do effortful things we have never done before. So
we do seem to have a central controller (one that cognitive science calls ‘working memory’), and
it will be a crucial character in this story. So the neural-networks revolution did not lead to the
demise of the central controller after all.

So one of the anti-intellectualist’s crucial conceptual tools (neural networks and their
respective associative mental processes) has been captured by the intellectualist. The new
intellectualism takes associations on board and is able to call itself ‘situated’ because of it. Moreover, it
seems able to explain all the fast, skilled, and low-effort phenomena that anti-intellectualism offered,
plus the cases of cognitive conflict that anti-intellectualists would be hard pressed to explain (given
their rejection of reflective processes). This is the birth of situated intellectualism.

2.2.2. A theoretical landscape of intentional action

It is now possible to depict a more complete theoretical landscape of intentional action (see Fig.
2.4). The main opposition is that between intellectualism (i.e. the view that all intentional actions
are produced by processes of reflection) and anti-intellectualism (i.e. the view that reflection is
not necessary for intentional action). All versions of intellectualism are representationalist, since
reflection is the manipulation of representations by a central processing unit. On the other hand,
anti-intellectualism can be representationalist (like the extended-mind theory) or anti-
representationalist (like radically embodied cognitive science). The claim that some intentional
actions do not depend on the reflective operations of a central executive is compatible with
representationalist and non-representationalist views of the mind.

Now, the label situated representationalism applies to all the views that posit associative
mental representations and a neural-network architecture. These are usually called Connectionist
models of the mind, and they are often contrasted with classical Cognitivist models, which I have
earlier called propositional intellectualism. But it is worth keeping in mind that connectionist models
admit of both representationalist and non-representationalist interpretations (Fodor & Pylyshyn
1988, 7–11). Gendler’s alief is a case of the former; Dreyfus’ skilful coping is a case of the latter.

---

33 The psychological notion of working memory, and its executive-control functions, are explained in §3.2.2 below.
A situated-representationalist view can fall into either of two camps: on the one hand, situated intellectualism holds that associative behaviours constitute intentional actions only insofar as they are controlled by higher-order reflective processes (like deliberation and planning). On the other hand, representationalist anti-intellectualism holds that the intentionality of associative behaviours is not always dependent on higher-order reflective processes.

All the views discussed in this section so far are representationalist. Anti-representationalist anti-intellectualism opposes all of them by holding not only that there are reflection-independent intentional actions, but moreover that they do not rely on representations that mediate between mind and world: rather, they are directed toward the world itself. This is the view presented by Merleau-Ponty (among other phenomenologists) and further developed by Dreyfus and the enactive approach.

---

34 Gendler endorses a form of situated intellectualism. Other recent situated intellectualist positions are those of Wu (2011, 2013, 2015), Fridland (2014, 2015), and Shepherd (2014). These views will be discussed in detail in Ch. 3–4.

35 Forms of representationalist anti-intellectualism have been defended by Clark (1997), and more recently by Brownstein & Madva (2012) and Brownstein (2014). See Chapters 3–4 for detailed discussion.

36 Anti-representationalist anti-intellectualism could be further subdivided according to the type of mental architecture endorsed, with Dreyfus being among those who endorse a connectionist architecture, and enactivism (Thompson 2007, 10–15) and Chemero’s (2009) radical embodied cognitive science appealing to dynamic systems theory rather than neural networks. Other recent positions along these lines are defended in Rietveld (2008), Klaassen et al. (2010), and Di Nucci (2013). See further discussion in Chapters 3–4 below.
The opposition between representationalist and anti-representationalist anti-intellectualism is sometimes expressed as the opposition between embodied cognitive science and radical embodied cognitive science, but, as we have seen, the term ‘embodied’ can also be appropriated by situated intellectualism. As seen above (§2.1.2), there’s debate within the anti-intellectualist camp: anti-representationalists criticize representationalists for not being radical enough; and representationalists (i.e. extended-mind theorists) question anti-representationalism’s capacity to provide a proper account of ‘representation-hungry’ cognitive activities like deliberation and planning. These discussions, although fascinating, can be bracketed while we examine the central issue to be investigated in what follows. For both radical and moderate anti-intellectualists can work as allies in the dispute against a common opponent: situated intellectualism.

Earlier I formulated this work’s main question as follows: What kind of agency and control do we humans really have over our behaviour, if most of it is automatic? How does it work, and what are its limits? Given that this is the main concern here, the representation debate (i.e. the debate about whether all intentional action implies representations) is secondary with respect to the intellectualism debate (i.e. the debate about whether all agentive control stems from reflective processes or not). I will thus focus on the debate between intellectualism and anti-intellectualism, specifically in relation to how they account for our control over automatic-associative behaviour.

2.2.3. How do we control associative behaviour (if at all)?

Intellectualism and anti-intellectualism differ with respect to the kind of intentionality they attribute to automatic-associative behaviour. Calling automatic behaviour intentional is from the outset controversial, even paradoxical, because the term ‘automatic’ is associated with mechanical, ballistic, and unintelligent processes. Despite this, both anti-intellectualists and intellectualists have an interest in explaining how automatic-associative processes can constitute intentional action, because automaticity is ubiquitous in human behaviour, and because it seems to be an essential component of habitual and skilful actions.

The intentional character of an automatic process can be defended in two main ways: by arguing that it was causally controlled by reflective motivational processes (like deliberation or planning), or by arguing that it exemplifies a non-reflective form of directedness to the world, like motor intentionality. In the following paragraphs I briefly summarize the main outlines and points of divergence between these two approaches, as they have been recently defended.

---

37 The issues concerning the nature of automaticity are examined in detail in Chapter 3.
It’s all about control: Deviant causal chains and how intellectualism became situated

Arguably, situated intellectualism had its origin in the search for a solution to one of the long standing problems of classical, propositional accounts of action: the problem of deviant causal chains. According to the classical cognitivist account, a behaviour B counts as an intentional action if it is properly caused by motivational mental states (e.g. desires, intentions) whose representational content includes B or is properly connected to B.\(^{38}\) Despite its advantages, defining an action in terms of its causal antecedents generates the problem that in some cases the right causes may cause the right behaviour in the wrong way, i.e. through a deviant causal chain.

To illustrate the problem, consider the classical story of a rock climber who would like to get rid of his fellow climber’s weight, and intends to do so by loosening his hold on the rope from which the latter hangs. His intention, however, makes him so nervous that he loosens the rope, thus getting rid of the other climber (Searle 1983, 83; cf. Davidson 1973). Consider also the story Sextus Empiricus tells about the painter Apelles, who was trying to represent the lather on a horse’s mouth, but, having failed too many times, “gave up, took the sponge on which he had been wiping off the colours from his brush, and flung it at the picture. And when it hit the picture, it produced a representation of the horse’s lather” (Outlines of Pyrrhonism [PH] XI.28). Similarly—Sextus adds—, the skeptic intends to attain tranquility by deciding between two incompatible but similarly powerful philosophical arguments; however, after much reflection he cannot decide between them, so he gives up. Then, after suspending his judgment, he finds the peace of mind he had been looking for.

Cases like those of the climber, Apelles, and the Pyrrhonian skeptic represent a problem for the intellectualist’s causal theory of action, because they show that an agent’s B-ing can be caused by the intention to B, and nevertheless fail to constitute an intentional action, since B can be produced by some strange deviation in the causal chain. What makes these causal chains deviant, it seems, is that at some point in time between forming the intention to B and actually B-ing an intermediate event occurs (the climber gets nervous, Apelles and the skeptic give up) that undermines the agent’s control over her behaviour, and the lack of control renders the subsequent

\(^{38}\) Donald Davidson’s account (1963), an especially influential theory of this kind, holds that the proper cause of an action is a primary reason, i.e. a motivation plus a belief. This primary reason is numerically identical to the neurophysiological event that causes the behaviour. Davidson’s account came at a point where the situation was that you could either rationalize an action by providing a non-causal explanation of it in intentional terms, or provide a causal explanation of it in purely physiological terms. This account’s great virtue was showing how you could have it both ways: primary reasons rationalize the action (since they are propositional mental states), and they also causally account for it (since they are token-identical to the neurophysiological event responsible for the behaviour). Providing a way to have the best of both worlds, the Davidsonian causal approach became—and arguably remains—action theory’s dominant paradigm.
behaviour unintentional (Schlosser 2007, 187–8). Hence, what the classical causal theory seems to lack is an account of control—more precisely, an account of how and in which cases reflective processes (like deliberation) and mental states (like belief, desire, and intention) properly control behaviour. Once this account of non-deviant causal control is ready, the problem of deviance will be solved, and the intellectualist causal theory will be able to fully account for intentional action.\footnote{Some intellectualists recently endorsing this view are Schlosser (2007, 188), Wu (2011, 58), Aguilar (2012, 9), and Shepherd (2014, 407).}

But let us take a step back. The problem of causal deviance seems to be an extreme case of a much broader issue, namely, the host of obstacles that reflection faces in successfully controlling behaviour. A deviant link in a causal chain is one such obstacle, but many other kinds of obstacles can be found in the psychological literature on automaticity and the processing limitations of reflective capacities.\footnote{As I argue below (§3.2.2), intellectual-reflective processes like deliberation and planning rely on working memory, an extremely limited cognitive resource. Given such resource limitation, it is unsurprising that much of our moment-to-moment sensorimotor behaviour occurs via automatic associations rather than propositional, rule-based processes. What is surprising is that the latter are able to control the former at all.} Everyday experience corroborates this: I may lose reflective control over my behaviour, e.g., when I intend to take a non-habitual route home when driving, but end up getting distracted and following the habitual path instead; or when I my mind wanders in the middle of a complex task (like reading or writing an essay) and I end up procrastinating. These are first hand, everyday experiences of the limitations of reflection, and of the many obstacles it faces in attempting to control behaviour.

Given the great complexity of the perceptual and practical worlds, and given our limited reflective capacities, the possibilities for reflection to lose control over our behaviour seem just as ubiquitous as automaticity itself. In order to stay relevant, a proper intellectualist account of action must show how, despite the countless obstacles, reflection is able to control moment-to-moment, automatic-associative behaviour. In other words, given the obstacles to reflective control (causal deviance included), intellectualists are forced to make situated intellectualists, i.e. action theorists that explain how limited reflective capacities can control association-based behaviour.

The situated intellectualist’s first move is a shift in focus from beliefs, desires and deliberation to intentions and planning. Intentions are a particularly useful mental state for the situated intellectualist because once formed, they resist being revised (unless an unexpected situation necessitates revision) and commit the agent to doing what she intends when the right time comes. So limited-capacity agents like us can employ intentions to coordinate our actions with other agents and with ourselves through time (cf. Bratman 1987). Moreover, when the time to act arrives, the agent mustn’t reproduce the deliberation process that led to form her
intention, because she is already committed; so through intentions past deliberations can structure the agent’s present and future behaviour, even in cases where the agent must act so quickly that she would not have time to deliberate upon what to do, thus minimizing the amount of practical reasoning required for intentional action (Bratman et al. 1988). In short, whereas beliefs and desires are merely potential conduct-influencers, intentions are conduct-controlling motivational states (Bratman 1987), and this is why they have been found to be useful in explaining reflection’s control over behaviour in associative, limited-capacity agents.

Further, intention and planning are tightly linked phenomena. A plan can be seen as an intention’s representational content (Mele 1992, 218), and in this sense plans can take a wide variety of forms, from very simple action representations (e.g. a non-conceptual, pictorial or sensorimotor sketch of myself performing a certain act) to a very elaborate set of instructions (a step-by-step propositional description of the process through which a complex goal is to be achieved) (Shepherd 2014, 397). Further, planning is a reflective process subjected to rational constraints (like consistency and a means-end hierarchical structure), which implies that behaviour performed in the execution of an intention can also be subjected to the scrutiny of practical rationality. Taken together, the broad conception of intention as a conduct-controlling representational state (whose representational content can be either conceptual or non-conceptual, either propositional or associative), and the integration between intentions and plans, empowers reflection to avoid several obstacles in controlling a broad range of behaviour, while limiting the reflective resources required for it to do so.

Now, how is it exactly that intentions can control associative and automatic processes? The details of how situated intellectualists relate intention to attention, perception, and motor behaviour, will be explored later on (§4.5). For now it should suffice to point to a particularly persuasive illustration. In Yarbus’ (1967) study of eye-saccades, participants were shown a painting (Ilya Repin’s An Unexpected Visitor) and then told to perform a series of tasks were while their eye movements were recorded. Yarbus found that the participants’ eye-saccade patterns varied systematically in accordance with the task performed (see Fig. 2.5 below). This lends support to

41 Modern, situated intellectualism adopts a broad conception of intention, that trespasses the propositional/associative and conceptual/nonconceptual boundaries (see e.g. Wu 2015, 4). Contrast this with e.g. Velleman’s (1985, 51) more classical view that intentions are beliefs that one will perform a certain action.
the intellectualist view that intentions can causally structure patterns of sensorimotor behaviour as automatic as eye saccades, and coordinate them with respect to a higher-order goal.\footnote{Wu (2015) and Fridland (2015) recently made this argument, earlier made by Rowlands (2006, chapter 11). The fact that eye saccades varied with respect to each task’s semantic content suggests that automatic processes are not only causally structurable, but also cognitively penetrable by higher-order reflective processes (i.e. that automatic processes are sensitive and responsive to the semantic properties of the agent’s occurrent intentions, and not merely causally dependent on them; cf. Fridland 2015).}

![Figure 2.5: Seven records of one subject’s eye movements (Yarbus 1967).](image)

What does this entail for the intentionality of automatic behaviour? It seems to suggest that associations are sensitive to norms, where the relevant norms are derived from higher-order reflective processes. In Gendler’s words (2008b, 573), the normative value of associations corresponds to the extent to which they are in line, or fall out of line, with reflective commitments. Associative normativity is thus intention-centred: an intentional associative behaviour’s success or failure is determined by the degree to which it conforms to the relevant intention’s representational...
content. Accordingly, the degree of control over associations would consist in the degree of approximation between the associative behaviour and the intention’s representational content.\(^{43}\)

In sum, having broadened the notion of intention and having shown that it is possible for reflective processes to inform associations, situated intellectualism holds that the intentional and normative character of associative states derives from the intentionality of reflective processes like calculation, deliberation and planning. Thus, automatic processes are under our control insofar as their performance coincides with the content of intentions and other higher-order mental states.

**No need to reflect: the anti-intellectualist reply**

An anti-intellectualist may question whether the Yarbus experiment illustrates the general nature of automatic-associative behaviour’s intentionality. After all, the situation in his study seems relevantly artificial: carrying out the instructions provided in the experiment (like estimating the family’s material circumstances, assessing how long the visitor had been gone for, or figuring out what the family was doing prior to his arrival) requires a level of intellectual engagement considerably greater than the one required by much of our moment-to-moment, habitual engagement with the world. It may be that intentions play such a controlling role in performing Yarbus’ tasks because they are so intellectually demanding. But the picture of automatic intentionality may turn out to be different if we focus on activities that are less focused on reflective and much more on sensorimotor processes.\(^{44}\)

This is the line along which Brownstein and Madva (2012) respond to situated intellectualism (specifically, to Gendler’s intellectualist theory of alief). In defence of anti-intellectualism, they hold that prototypical associative states have something relevantly similar to Dreyfus’ Merleau-Ponty-inspired notion of ‘felt tension’. Gendler holds that alief includes an affective component; Brownstein and Madva agree and hold that this component should be understood as directing the agent toward a certain alleviating motor compensation without any recourse to reflection. Take Merleau-Ponty’s museumgoer example:

> The museumgoer moves away from the large painting to get the best view, without ever having to consider that her initial view was flawed. Of course, she may reflect on where to stand and judge that her automatic impulse is misguided. But absent any defeating reflective considerations, her automatic impulse drives her to act at the right time in the

\(^{43}\) This is the view recently endorsed by Shepherd: “the exercise of control is essentially a matter of an agent’s bringing behavior to match the content of a relevant intention (or more broadly, a relevant motivational state)” (2014, 398).

\(^{44}\) Cf. Rowlands (2006, 204).
right way, just by virtue of being appropriately responsive to the array of subtle environmental features. (2012, 420)

Gendler’s account of the structure of alief should be accordingly modified from R-A-B (i.e. composed of Representational, Affective, and Behavioural components) to F-T-B-A: associations present a Feature of the world as loaded with a certain affective Tension; this tension solicits a Behaviour that is directed towards Alleviating said tension. Crucially, the felt tension generates a dynamic feedback loop between behaviour and alleviation, in which the search for further alleviation leads the agent to perform subsequent behaviours, continuously tending toward further alleviation while guided not by a clear representation of the goal, but by the felt modifications to the sensation of tension. Brownstein and Madva’s dynamic account of automatic-associative behaviour thus resembles Merleau-Ponty’s motor intentionality, in that automatic behaviour constitutes a directed process, but the directedness is not based on either a well-defined representation (e.g. of the specific place to be reached), or on a previous reflective process (e.g. a calculation of the specific distance to move away from the painting).[45]

Nor should the success or failure of associative behaviour be judged by reference to the representational content of a causally prior mental state, since the dynamic feedback between the association’s behaviour and alleviation displays a normativity of its own. This is evident, according to Brownstein and Madva, in that aliefs are capable of error (the museumgoer may overcompensate by going too far from the painting) and self-modification (she would then seek to alleviate tension by moving a little closer). Such normative processes are guided only by the feeling of tension, and not by prior reflection (Brownstein & Madva 2012, 420–5). The normativity of automatic-associative behaviour is therefore fundamentally independent from the reflective normativity of beliefs, intentions, and plans. Thus the anti-intellectualist replies to the situated intellectualist by arguing that the intentionality of everyday, habitual associations cannot be subsumed into act intentionality, but must rather be considered an autonomous operative-motor intentionality, that works independently from the reflective operations of a central controller.

[45] Brownstein and Madva’s anti-intellectualism is thus representationalist, since they consider the association’s Feature element to be representational. Rietveld (2008) defends a similar, albeit anti-representationalist, anti-intellectualism, holding that the normativity of “unreflective action” has its foundation in the phenomenon of “directed discontent” (a term borrowed from Wittgenstein, analogous to Brownstein and Madva’s ‘felt tension’), i.e. an affective reaction proper to skilled agents, whereby they “see what should be done to improve the current situation or solve the problem; to perceive and act on possibilities for action” (2008, 980). Rietveld and his colleagues have suggested (Klaassen et al. 2010) that directed discontent, which occurs in the narrative time-scale, has its roots in the enactive notion of valence, understood as the originary process of directed affect which occurs in the micro-scale of milliseconds.
§2.3. Conclusion

This chapter has attempted to trace the recent history of discussions about the nature of intentional action, and then focused on the question of how we should account for the control we exert over automatic-associative processes. I have just argued that there are two incompatible answers to this question coming from the intellectualist and anti-intellectualist camps. Which one is preferable?

On the one hand, the anti-intellectualist argues for the independence of automatic-associative normativity from reflective processes. This answer implies that automatic-associative processes display a normativity of their own. But do they? Automatic processes have often been conceived as reflex-like neural firings that blindly run to completion. So is there a reflection-independent normativity of automaticity? If so, then what properties does it have? Chapter 3 tackles these questions.

On the other hand, intellectualists claim that the control displayed even in fine-grained, skilled bodily actions is ultimately caused by reflective processes like planning and deliberation. This seems unlikely, as skilful coping has traditionally been anti-intellectualism’s stronghold. Can situated intellectualism convincingly argue that even the most expert kinds of skilled activity (which are usually considered to be too context-sensitive and complex to be a matter of reflection) require reflection? Answering this question is the task of Chapter 4.

Now, even if the intellectualist argument succeeds in this respect, anti-intellectualism has another, currently more overlooked, argument route. I.e. she can argue that the intellectualist control based on reflective processes depends on more fundamental, pre-reflective forms of sense-making that constrain and inform reflective control. If this is so, agentive control would depend on the non-reflective forms of control that allow for such originary forms of sense-making. Chapter 5 will explore an argument along these lines based on the phenomena of framing effects.
Chapter 3: What is automaticity?

Last chapter provided a theoretical landscape of intentional action, and argued that the question of control (i.e. What kind of agency and control do we humans really have over our behaviour, if most of it is automatic? How does it work, and what are its limits?) is a central challenge that must be faced in order to advance toward a successful theory of human agency. This chapter seeks to advance toward an answer by discussing the nature of automatic-associative processes. These are often interpreted as blind—i.e. largely mechanical, ballistic, and unconscious—phenomena, and such interpretation makes the question of control more puzzling, since it is unclear how we can exert agentive control over blind, unintelligent processes. This is a relevant issue for the field of possible answers to the question of control, because if automatic processes are blind, there would be no way for us to control them ‘from within’, as anti-intellectualists hold, and all control would depend on our executive power to structure them top-down via reflective processes.

In this chapter I will argue against the blind interpretation of automaticity. Evidence shows that (at least some) automatic processes are highly context-sensitive, display normative properties, and are consciously accessible in relevant respects. If this is so, then the path is open for accounts of control other than the intellectualist’s top-down model. (To clarify, I do not intend to make claims about automatic processes as a whole: this may not even be possible, given the huge variety of kinds, structures, and functions that automatic-associative mental processes have. Rather, my point is that at least some automatic processes can be highly context-sensitive, responsive to norms, and consciously accessible; and therefore that there is nothing in the nature of automaticity that impedes this.)

I will (§3.1) introduce the discussion by means of a paradox: automaticity and control seem to be mutually exclusive, but at the same time automaticity seems necessary for a great number of cases of control. The paradox of automaticity informs the discussion in this Chapter and the next. I will then (§3.2) introduce the overarching theoretical framework within which studies on automaticity have taken place: the dual-process framework of the mind. This will make it possible to give the precise technical sense to crucial terms, like ‘automaticity’, ‘intuition, and ‘reflection’, that I have been using until now appealing merely to their commonsense meanings. Equipped with these rich conceptual tools, I move on to build (§3.3) a critical assessment of the blind interpretation of automatic-associative processes. After defending the claim that a normative and intentional understanding of automaticity does more justice to what we know about it, (§3.4) I explain what this reveals about automatic normativity.
3.1. The paradox of automaticity

There has been a tendency in both philosophy and psychology to consider automaticity and control as opposites: calling a mental process ‘automatic’ has often been taken to imply that it cannot be stopped, altered, or corrected, and often that it is unconscious; all of which suggests lack of control. And vice-versa, a process that the agent is able to stop, revise, correct, or be conscious of, has often been taken to be controlled and therefore not automatic. Yet there is also the view that automaticity is necessary for control: in order for us to perform any skilled activity, from tying our shoelaces and walking to playing sports and doing complex math, we must have automated the action’s basic routines in order for a higher-order coordination of activities to be possible. This generates a paradox, since automaticity seems both incompatible with, and necessary for, control. Let us look at the two sides of the paradox in a bit more detail.

3.1.1. Why automaticity threatens control

John always rides his bike through the same path as he goes back home from his office. One day, as he gets ready to leave work, he reminds himself that his son’s birthday is coming soon, and he has to take a detour and pick up his order at the candy store. He remembers his plan again as he finds himself in front of his door, having already taken the same old route. Slips like that are very common, and suggest that automaticity threatens our control over behaviour, insofar as it can make us act contrary to our governing intentions (Amaya 2013). But slips are just one of the many challenges that automatic processes pose to agentive control if, as numerous and solid research programmes suggest, automaticity is a pervasive feature of everyday life.

From argument evaluation, judgment and decision-making, to imitation, category activation, and goal activation, automaticity seems to exert influences on our behaviour that are independent from our intentions, often go unnoticed, and therefore constitute a threat to intentional control. And even the production of intentional action seems to require automatic cognitive or sensorimotor sub-routines: a handshake requires a complex activation of dozens of muscles, none of which we could directly control. This line of thought reveals the full scope of automaticity’s threat: if, on the one hand, automaticity (being unstoppable, unmodifiable, and

---

1 As will be described below, the picture of the relationship between automaticity and control has become much more complicated than a duality of opposites. However, the original description of the concepts (Schneider & Shiffrin 1977; Shiffrin & Schneider 1977) still bears influence on contemporary research, and is therefore worth taking as a starting point for conceptual analysis. (For a recent proposal in this spirit see Wu 2013a.)

2 This is a widespread view in the psychological literature. For an influential account see Logan (1985).

3 For illustrations of these types of automatic behaviour, and how they seem to threaten control, see the Introduction above.
more often than not unconscious) implies the absence of control, and on the other, automatic routines are pervasive in every aspect of our behaviour—including intentional action—, then the realm of control seems to be minimal, if not altogether nonexistent.

3.1.2. How automaticity underlies control

Consider automaticity now from a different perspective: that of skill acquisition. Novices in any skill (from riding a bicycle to playing an accordion to speaking in a second language) require a taxing amount of reflective control in order to perform the basic tasks. But through continued exposure, the simple tasks start to become easier and less cognitively demanding. And as the simple tasks become automated, the agent’s horizon of possible activities expands to include more complex behavioural sequences.

This suggests that an increase of automaticity not only does not make control disappear, but actually *increases* it. In any activity requiring skill, once the simpler routines become automatic the agent is able to tie them together in threads of more complex fluid action. It would be impossible for an average human to play a whole piano piece while reflectively controlling each finger movement, or to have a fluid conversation while inferring each sentence’s proper construction from the rules of grammar. This shows just how necessary automaticity is for everyday, habitual control. Once a skill is acquired, skilled behaviour is no longer the product of rule-based calculation or deliberative analysis (there are just too many variables and too little time to manipulate them), but rather of association-based pattern recognition, fast intuitive judgment, and automatic sensorimotor activity (Dreyfus 2002). Intuitive judgment allows for faster and more accurate identification of relevant information, and quicker decision-making.

This applies to skilled performers of all kinds, from typists, firefighter commanders, and cricket players to chess players and fingerprint examiners.⁴

In sum, automaticity seems to threaten our control over everyday activities by making them rigid, intention-insensitive, and susceptible to environmental influences that seem to remain

---

⁴ For typists see Liu et al. (2010) and Logan & Crump (2010); for firefighters, Klein et al. (1986); for cricket players, Land & McLeod (2000); for chess players, Chase & Simon (1973); for fingerprint examiners, Thompson (2014). General psychological discussions on the links between automaticity, habit, skill, and control can be found e.g. in Bryan & Harter (1899), LaBerge & Samuels (1974), Logan (1985), Abrahamse et al. (2013). The philosophical debate surrounding this issue will be explored in Chapter 4. For now, it is worth remarking that expert judgment is *not necessarily superior* to slow, reflective judgment. While often astonishingly fast and accurate, intuitive expert judgment is sometimes notoriously off-mark, because it works on the expectation that the associations acquired through practice are relevant to the situation at hand. When they are not—perhaps because a relevant aspect of the decision-making environment has changed—, intuitions and expert judgments can misfire and fall prey to cognitive biases (Morewedge & Kahneman 2010).
outside of our awareness; but it also seems to be necessary for the control of any minimally complex action—a class within which arguably most human action falls.

How can we make sense of this? In order to dispel the air of paradox I will take two steps: the first one, which I take in this chapter, is to clarify the nature of automaticity, its relationship to normativity and to intentionality. The second one, which will be next chapter’s topic, is to clarify the nature of our control of skilled actions. But before delving into the concept of automaticity, it is important to specify the theoretical context within which it has developed, i.e. dual-process research programs.
§3.2. The dual-process framework

In this section (2.1) I trace the modern history of dual-process models up to a point where the crucial distinction between two process types is defined by reference to the concept of working memory. This motivates (2.2) an examination of the notion of working memory. I finish by (2.3) discussing some of this literature’s implications for the notion of control.

3.2.1. Some history

Conceptions of human cognition that split the mind in different parts, systems, or process types have a long history, going at least as far back as Plato and Aristotle in the West. This approach to cognition has enjoyed a revival in various disciplines, including those studying perception, learning, memory, reasoning, decision making, attention, emotions, addiction, social cognition, economic behaviour, moral judgment, and a host of other domains. Since the 1970’s, countless dual-process models have been put forward to account for diverse psychological phenomena. Around that time, researchers working on different research programmes independently found evidence for the existence of tensions between cognitive processes, and expressed their findings using different conceptual oppositions to characterize the multiplicity of conflicting processes observed (automatic/controlled, fast/slow, effortless/effortful, implicit/explicit, parallel/serial, domain-specific/domain-general, associative/rule-based, evolutionarily-old/evolutionarily-modern, voluntary/involuntary… the list goes on). Later, similarities between these approaches became apparent, and a trend toward the unification and systematization of dual-process models took place. The result was a broad appeal to two systems in charge of different cognitive tasks. In these early synthetic models, ‘System 1’ is in charge of fast, automatic, effortless, unconscious, heuristic, associative, working-memory-independent processes; and ‘System 2’ is responsible for controlled, effortful, conscious, analytic, working-memory-dependent processes.6

The two-systems picture was later deemed too simple to account for the complexity of studied phenomena (Stanovich 2004, Evans 2010a). A crucial reason for this critical verdict is that the label ‘System 1’ encompasses not just one, but actually countless subsystems, each with its own evolutionary history, physiology, and functionality, thus giving the misleading idea that

5 For more on the not-so-recent history of the dual-process approach in the Western tradition, see Frankish & Evans (2009). For accounts of the dual-process framework’s recent history and evolution, see Evans (2008) and Evans & Stanovich (2013). For a thorough historical review of the notion of automaticity within the dual-process context, see Moors & De Houwer (2006; 2007).

6 Synthetic dual-system models were developed e.g. by Stanovich & West (1999) and Kahneman & Frederick (2002). Kahneman (2011) employs the two-systems vocabulary, but acknowledges its limitations, which will be highlighted below.
there is some common, unified nature for what actually turn out to be very different, often independent processes. It became clear that a single system cannot be responsible for so many functions (including perception, implicit learning, intuitive judgement, and implicit social cognition), and likewise, that just one other system is in charge of a host of others (like the executive control of attention, explicit learning, the construction of mental models, inferential reasoning, planning, consequentialist decision-making, etc.). Given this, the most sophisticated current versions of the model reject the use of two-systems talk, and defend a different dual categorization, based on the idea that cognitive processes can be split into two main types. A consensus seems to be forming around the view that processes of one of these types (‘Type 1’ or ‘intuitive’ processes) is defined by its independence from working memory, while processes of the other type (‘Type 2’ or ‘reflective’ processes) require the use of working memory. Reflection is a slow, sequential, and limited-capacity process because working memory is a limited resource whose use implies cognitive effort. Likewise, intuition is fast, effortless, and able to work in parallel because it is working-memory independent.\footnote{7 For further discussion see Samuels (2009), Frankish (2010), Evans (2010a; 2010b), Evans & Stanovich (2013). Not everyone is on board with employing working memory to define reflection and intuition, however. For a critical approach, see Carruthers (2012).}

So far I have constantly used the concepts of ‘reflection’ and ‘intuition’, but only intuitively, without properly defining them. I will adopt the proposed dual-process convention and understand ‘reflection’ as referring to Type-2, working-memory-dependent cognitive processes, and ‘intuition’ to Type-1, working-memory-independent processes. The notion of ‘working memory’, being crucial to understanding these notions, is worth exploring in some detail.

### 3.2.2. Working memory

**The concept and its origin**

The term ‘working memory’ stemmed from research on short-term memory. The latter was originally conceived as the capacity to hold a few items in mind for a brief time, the ‘store’ where you, e.g., place a phone number in order to dial it. The digits that make up the number are stored in short-term memory from the moment you hear them until the moment you are done dialling, and after dialling short-term memory can then be wiped out and used to store other bits of information. It actually gets wiped out rather quickly, because it has a very limited storage capacity, and we require it for other tasks.

But researchers noticed that performing tasks of this kind involved much more than just temporarily storing information. For one, when we do things like holding a phone number in
mind we often rehearse the number in inner speech to keep them, and their order, retrievable, and often lose it when we don’t rehearse it. Further, we must effortfully focus our attention on the numbers and avoid getting distracted away from them, while also dedicating some of our attention to concurrent tasks (like picking up and manipulating the phone). So what was previously conceived as nothing but a storage system that gets continuously wiped out, has turned out to be a more complex and interesting faculty, whose tasks include selecting the information that is relevant to a task’s performance (hence the ‘working’ in ‘working memory’), and then maintaining this information accessible by means of the executive control of attention, which often requires inner-speech rehearsal or the construction of mental models. Working memory is nowadays conceived as the complex, limited-capacity faculty in charge of all these operations.

Working memory, attention, and cognitive control

As with short-term memory, researchers have found that there are differences between individuals’ working-memory capacities. These are traditionally measured by the number of items a person can recall while performing an unrelated, attention-demanding task. However, rather than consisting in a certain amount of items a person can maintain in memory, working-memory capacity seems to measure the individual’s ability to maintain attention focused on a task despite distracting and cognitively demanding forces. This ability is what in turn accounts for the number of items an individual can maintain in an accessible state. In short, working memory capacity seems to directly measure a person’s ability to control attention to avoid distraction from interfering information, and measures memory storage only indirectly (Engle 2002). The cognitive control of attention implies the ability to inhibit distracting forces, build

---

8 The term ‘working memory’ appears to have been popularized when Atkinson & Shiffrin (1968) used it to express the goal-oriented aspects of short-term storage. Baddeley & Hitch (1974) then developed the “multi-component” model, which has now become the classic referent, but is not the only game in town, as will be discussed further below. (For more on the concept’s development and the current discussion, see Baddeley (2007; 2012). Evans (2008; 2010b) explains the centrality of working memory for the dual-process framework.)

There is debate about whether working memory can directly access propositional attitudes, or only sensory-based contents (like perceptions, memories, affects, mental imagery, or inner speech) (cf. Carruthers 2009; 2011). This debate has serious consequences, among other things, for our conception of self-knowledge (i.e. whether we know our own beliefs and desires directly, or through interpretative processes like those we use to know other people’s minds). However, I shall leave these discussions aside, since they are tangential to the crucial questions about the role working memory—and hence reflection—plays in agentive control, and about the relationship between control and automaticity. Either through propositions or through sensory-based mental models and inner speech, working memory is crucial for the executive organization of behaviour in the service of a higher-order goal.

9 For example, one way of measuring working memory capacity is presenting people with a sentence and an unrelated word, and asking them to judge whether the sentence is true or false while remembering the subsequent word. Working-memory capacity is quantified as the amount of unrelated words the person can remember while evaluating the sentences (Engle 2002).
and rehearse inner models of the information in order for it not to fade from memory, and efficiently divide attention between simultaneous tasks.

Working memory underpins all higher-order cognitive tasks, like hypothetical thinking, reading comprehension and writing, reasoning, planning, critical argumentation, and also the capacities for self-control in the face of temptation, for persistence in the face of obstacles, and for empathy as a product of looking at a situation from another person’s perspective. All these activities require the use of a general-purpose capacity that receives the outcomes of specialized (e.g. sensory, emotional, motor, or imagistic) sub-processes, and is able to attentively select, retain, and manipulate their outputs in the service of a higher-order, personal-level, consciously accessible goal. Since working memory capacity is a limited resource, the processes that require it—i.e. reflective or Type 2 processes—tend to be effortful, slow, and serial (since they must process information in the right order to produce the expected outcome, e.g. in a complicated multiplication). This is in contrast with the effortless, fast, and parallel traits that characterize intuitive, Type 1, associative processes that work with full independence from working memory. The flipside of this is that Type 1 processes tend to be very domain-specific and work in the same way every time they are activated, while Type 2 processes are domain general and can yield different results each time they work. It is clear from this that Type 2 processes require the products of Type 1 processes as inputs for working-memory-demanding tasks. Working memory has the power to coordinate Type 1 processes by focusing attention on them, thus making them active, and then manipulating their outputs in order to produce a desired, task-related outcome.

The origins of working memory

Now, if automatic processes are so fast and efficient, why do we need slow and effortful working memory for? Why are the myriad associative, experience-based sub-systems insufficient for successful cognition and action? From an evolutionary perspective, it does in fact make sense for the human mind to be a conglomerate of cognitive subsystems, since each one of these could have been separately modified by natural selection, and thus become adapted through its own semi-independent evolutionary history to perform specific cognitive functions.\footnote{For conceptions of the mind along these lines, see e.g. Simon (1962), Dennett (1991), and Carruthers (2006).}

However, if the mind were entirely made up of domain-specific cognitive sub-systems, whose outputs are mostly independent from one another’s, then the existence of cognitive activities like deliberation and planning would be a mystery, since they require the synthesis of multi-source information and its coordination toward unified behavioural results. In order for
those synthetic functions to take place, the mind would require some sort of ‘global workspace’ in which the results of its various subsystems could be posted and used as inputs for further processing. And tasks such as planning and deliberation would require that within this global-workspace architecture there be a faculty that selects, retains and manipulates the products of Type-1 process, to coordinate them in the service of a goal. This is working memory.\(^\text{11}\)

**Two conceptions of working memory**

So far I have presented a general picture of what working memory is and how it works. There is much debate about the details of this picture, however. In fact, theorists disagree about working memory’s specific function. Since this disagreement affects what counts as reflection and what does not, it is important to mention it here. The disagreement is about two ways to describe the function of working memory: a broad one and a narrow one.

The broad view, which I have depicted above, holds that working memory has a plurality of functions that include selecting, retaining, and manipulating goal-relevant information. These functions are performed in large part through the executive, top-down control of attention. In fact, in the broad view executive attention and working memory are so intimately related that they sometimes get described as largely equivalent.\(^\text{12}\) There is also a narrower view, according to which working memory’s specific role is just to manipulate the relevant information. The tasks of selecting and maintaining such information correspond to other so-called executive functions: **inhibitory control** (the ability to direct attention so as to inhibit predominant tendencies and keep focus on certain goal-relevant objects) and **cognitive flexibility** (the capacity to shift between tasks and sets of rules in cognitive processes).\(^\text{13}\)

Why is this disagreement relevant for the present discussion? Because many processes that would count as reflective in the broad view would not in the narrow view. Examples are

---

\(^\text{11}\) For discussions of the ‘global workspace’ theory and its relationship to working memory, see Baars (1997, 39–42), Carruthers (2013) and Wu (2013b).

\(^\text{12}\) See Engle (2002). Baddeley himself holds that “[t]he capacity to direct and focus attention is perhaps the most crucial feature of working memory” (2007). In his multi-component model, working memory is divided into a **central executive**, whose key functions are the executive control of attention (i.e. the direction, division, and shifting of focus), and a number of slave sub-systems (the **phonological loop** and the **visuospatial sketchpad**) that maintain and rehearse sensory information under the executive’s direction. Evans (2008; 2010a; 2010b; Evans & Stanovich 2013) is largely following Baddeley’s model when he claims that working-memory-dependence is the crucial distinction between Type 1 and Type 2 processes. So all these authors endorse what I call the broad view.

\(^\text{13}\) See Diamond (2013) for a review of the concepts, the discussion about its inter-relations, and an account sympathetic to the narrow view. See also Mercier & Sperber (2009) for a philosophical justification of the narrow view, and Carruthers (2012) for criticism of the broad view. In her philosophical discussion, Nagel (2014) constructs the concept of working memory, and consequently that of reflection, in accordance with the narrow view.
understanding a complex sentence, in which the first parts must be kept in mind until the end (e.g. Ships are believed, and with some considerable justification, to often be lived on by sailors), or looking for a friend’s face in a crowd. These activities require the effortful and selective control of attention, but they do not imply a manipulation of information, i.e. they do not imply that intuitive outputs are used as inputs for further processing rounds, which ultimately lead to the production of a mental model, an attitude, a reasoned inference, or a plan (Nagel 2014, 229). Hence theorists endorsing the narrow view consider that activities like those just mentioned, although cognitively effortful, do not deserve the name ‘reflection’.

And the narrow theorists are right, if we take ‘reflection’ to be synonymous with ‘reasoning’. Although this use of the terms is useful in some circumstances, the general view (in which the term ‘working memory’ and ‘reflection’ are largely coextensive with the terms ‘executive function’, ‘executive attention’ and ‘cognitive control’) is much more useful for the purposes of this work, for two main reasons: first, the theorists that construct dual-process theory through the concept of working memory (crucially, Evans and Stanovich) use the broad conception of working memory to distinguish reflection from intuition. Second, and perhaps most importantly, the broad view allows us to tie dual-process theory with intellectualism about action via the concept of reflection.

3.2.3. Reflection and control

I have defined intellectualism (§2.1) as the view that reflection is necessary for the production of intentional action. Now that the notion of ‘reflection’ has received its proper, technical definition, intellectualism can itself be more clearly specified. Situated intellectualists hold that intentional action requires that behaviour be produced in the service of an intention, construed broadly (§2.2) as any mental state (propositional or non-propositional, conceptual or non-conceptual) whose content is the representation of the intended state of affairs. Reflection is required for the production of intentional action because structuring cognitive or sensorimotor behaviour in the service of a goal requires the use of top-down attention to select the goal-relevant information, maintain it accessible, and often manipulate it in order to construct plans or mental models of how to achieve the intended goal. and since all of these cognitive processes

14 The sentence is an example taken from Baddeley (2007) and used by Carruthers (2012). The crowd example is taken from Mercier & Sperber (2009) and used by Nagel (2014).
require working memory (broadly understood), the production of goal-oriented behaviour requires the use of working memory.\footnote{Recent intellectualist accounts, that see the executive control of attention as necessary for intentional action production, include those of Annas (2011), Sutton et al. (2011), Wu (2011; 2013b; 2015), Shepherd (2014; 2015), and Fridland (2014; 2015). These will be discussed in detail in Chapters 4 and 5.}

Now—one may want to ask the intellectualist—, why is reflection necessary for intentional action? Why, that is, are intuitive processes not sufficient? After all, we seem to go about intuitively through a large portion of our lives, particularly when we engage in habitual, yet purposive, behaviour (like brushing our teeth, walking down the street, moving our bodies through the soccer field in a game, reading someone’s body language…).

To defend her position, the intellectualist must hold that no purely intuitive behaviour counts as a full-blown intentional action. There are two main strategies the intellectualist may use defend this view. One of them (which I will call the \textit{blindness strategy}) is to deny that intuitive, i.e. automatic-associative processes have any sort of intentionality: they are blind and purely reactive to environmental forces, so reflection is necessary to elevate unintelligent reactions to the level of intelligent action by threading them together in the service of an intention. The other way to defend intellectualism (which I will call the \textit{insufficiency strategy}) is to argue that, even though automaticity may display a certain kind of intelligence, intentionality, or normativity, such properties are nevertheless insufficient for the production of full-blown intentional action because purely intuitive behaviour lacks some crucial feature. In the following section I will critically assess the blindness strategy by assessing the blind interpretation of automaticity. I leave the insufficiency strategy for Chapter 4, which examines the nature of control in skilled action.

To summarize: I have (§3.1) presented the conceptual difficulties surrounding the notions of automaticity and control through the paradox of automaticity; I have then (§3.2) provided a precise definition of some of the key terms involved—automaticity, intuition, and reflection—with the help of the dual-process framework and research on working memory. In the chapter’s last, and crucial, section I will explore empirical evidence about automatic processes in order to assess the intellectualist’s blindness strategy. Are automatic processes really blind, unintelligent, and merely reactive? Or does research into them find evidence for some sort of automatic intentionality?
§3.3. Are automatic processes blind?

The blind conception of automatic-associative processes has a long history. Perhaps the strongest evidence for the uncontrollability of automaticity comes from Stroop-type effects (Logan 1985, 378–379), where unattended dimensions of a stimulus interfere with the attention-demanding task that the subject attempts to perform.\(^\text{16}\) Findings like these give strength to the view that automatic processes are *mechanical* (the relevant input triggers them almost invariably), *ballistic* (once activated, they run to completion), and *unconscious* (their activation, content, or effects on subsequent behaviour are for the most part introspectively unavailable). The blind interpretation of automaticity is the view that automatic processes are mechanical, ballistic and unconscious; and its upshot is the claim that automatic processes are not in themselves controlled or controllable. In fact, control seems to imply sensitivity to reasons or norms, the capacity for error-detection and -correction, and awareness of the controlled process. Automatic processes, being blind, lack all of that.

In this section I gather evidence that challenges the blind interpretation. If automatic processes cannot be so interpreted, the negative consequence would be that the intellectualist’s blindness strategy is blocked, and the positive consequence is that anti-intellectualism is reintroduced as a relevant alternative for the explanation of intentional action. In what follows I examine whether automatic processes are (3.3.1) mechanical, (3.3.2) ballistic, or (3.3.3) unconscious.

### 3.3.1. Are automatic processes mechanical?

Automatic processes have been understood to be reflex-like, activated by the mere presence of a given input. This view has been present from the beginning of the concept’s history in modern

---

\(^{16}\) In the Stroop test (Stroop 1935, MacLeod 1991), participants are instructed to name the colours of the inks in which some symbols are printed. The symbols are themselves colour names, and ink colours vary, sometimes matching the printed name (e.g. “RED BLUE YELLOW”), and some others not entirely matching it (e.g. “RED BLUE YELLOW”). By measuring the speed at which participants identify the colours, psychologists discovered that it takes people significantly longer to correctly report the ink colour when it is incongruent with the printed word. They inferred from this that participants automatically read the printed word (since reading had become automatic due to overpractice), and in the incongruent setting they had to use reflection to override the reading process’ output in order to provide the correct one. This effect is not restricted to words and colours: you can have different congruent-incongruent settings and still get Stroop effects. Because in these cases the automatic process is antagonistic to the intended activity, the automatic process’ influence on performance can be measured via the increase in response time it causes.
psychology, and carries its influence to this day. However, no sensible account, no matter how mechanistic, would deny the power of context. After all, any mechanical interaction between two bodies yields different outcomes depending on whether they meet in the air, under the water, or in outer space. Similarly, mechanistic interpretations of automaticity view the stimulus’ impact on the organism to be conditional on the context. In fact, Stroop effects themselves have been found to be modulated by contextual features like how many dissonantly coloured characters each word contains, the participants’ direction of attention, and the goals of the the participant’s task. This goes against the traditional orthodoxy that automatic processes are attention-independent, and suggests that the context that determines an association’s activation includes not only external features of the environment, but also the agent’s own cognitive dispositional and occurrent states at the time of stimulus reception.

But if automatic process activation is so sensitive, does it make sense to call it mechanical? In fact, the broader and more widely encompassing the conditionality of automaticity is, the farther away it seems to stray from a mechanistic conception. In fact, there should be a point where automaticity is so widely conditional—its functioning so sensitive to the agent’s particular practical situation—that calling it ‘mechanical’ would no longer do it justice, since the ‘triggering stimulus’ would turn out to be a specific practical moment, whose specification includes the agent’s own cognitive and motivational states. It would seem like automaticity responds to a ‘holistic sense of context’, rather than an isolatable, unitary stimulus. Cases 1–2 show that, contrary to the mechanical interpretation, automatic process activation is not unaffected by “where the current focus of conscious attention is, what the individual was recently thinking, or what the individual’s current intentions or goals are” (Bargh 1997, 3).

17 Schneider and Shiffrin defined an automatic process as an activation of a series of neural nodes in which “[t]he sequence of nodes (nearly) always becomes active in response to a particular input configuration, where the inputs may be externally or internally generated and include the general situational context” (1977, 2). The view has remained influential: “The essence of TASS [=intuitive] subprocesses is that they trigger whenever their appropriate stimuli are detected, that they cannot be selectively ‘turned off’ ” (Stanovich 2004, 52). In a review of literature on automaticity in social behaviour, Bargh et al. (1996, 252) find that “[r]ecent research has shown that attitudes and other affective reactions can be triggered automatically by the mere presence of relevant objects and events, so that evaluation and emotion join perception in the realm of direct, unmediated psychological effects of the environment.”—a view that Gendler (2008a, 644) quotes approvingly. (Cf. also Antony 2002; Levy 2008, 2.) Fridland (2015) critically reviews accounts that equate the notions of ‘automatic’ and ‘unintelligent’.

18 Studies demonstrating the context-dependence of Stroop effects include Francolini & Egeth (1980), Kahneman & Henik (1981), Tzelgov et al. (1997), and Besner & Stolz (1999). These studies contributed to the rejection of the traditional view that automaticity is attention-independent, which had been proposed in LaBerge & Samuels (1974), Posner & Snyder (1975), Shiffrin & Schneider (1977), Schneider & Shiffrin (1977), and Hasher & Zacks (1979). The context-dependent conditionality of automatic process activation is supported by a wide range of social-psychological studies; see Bargh (1989), Blair (2002), and Casper et al. (2010).
Case 1: Automatic process activation depends on motivation

Richeson & Ambady (2003) instructed White female participants to cooperate with another person in a task, either in the role of superiors (i.e. they had to evaluate the partner’s performance) or as inferiors (i.e. their partner would evaluate theirs); they then found out that their interaction partner was a Black woman. By means of an Implicit Association Task\textsuperscript{19} it was found that those who were assigned a subordinate role displayed significantly lower automatic racial bias activation than those who were in a superior position.

In another study (Van Bavel & Cunningham 2009, experiment 2), participants were asked to memorize the names and faces of some unfamiliar people—including some Black people—as members of their team, and the faces of some others as members of an opponent team they would be competing against. By means of a Response-Window Priming Task\textsuperscript{20} the researchers assessed whether racial bias activation was affected by their self-categorization as members of this brand-new, totally arbitrary, mixed-race group. Results showed that participants revealed the standard activation of negative racial bias toward Black people in the opposite team, and even toward unknown people outside of their team, but displayed a positive bias toward all members of their team, both Black and White. Researchers concluded that the categorization of one’s self within a novel, arbitrary group modulates the automatic activation of social categories. These results, and a host of similar others (cf. Blair 2002, 244–248), show that motivational processes stemming from the adoption of a social role modulate the activation of automatic processes.

Case 2: Automatic process activation depends on the direction of attention

The automatic categorization processes that are elicited by one and the same stimulus also depend on the task one is set to perform. Macrae et al. (1997) asked participants to watch a series of photographs (some of women and some of household items), but their tasks varied: a third of the participants had to identify whether there was a white dot in the image; another third had to identify whether the depicted object was animate or inanimate; and the latter third had to simply report that an image had become

\textsuperscript{19} In the Implicit Association Task [IAT] (Greenwald et al. 1998), participants are presented with a word or image that they must classify into one of two sets of categories as quickly as they can while making as few mistakes as possible. By contrasting the relative latency between responses, the test measures how strongly a category from a given pair (e.g. Black or White) is implicitly associated to another category from another pair (e.g. Pleasant or Unpleasant). When the test finds a strong association between a social-group category and an evaluative category, then it has found evidence of an implicit bias. (For example, the test finds evidence for implicit pro-White bias when people perform more quickly while categorizing pleasant words and Whites (and unpleasant words and Blacks) together, in contrast with categorizing unpleasant words and Whites (and pleasant words and Blacks) together (Rudman 2004).) Harvard’s Project Implicit website (https://implicit.harvard.edu) provides a hands-on IAT experience.

\textsuperscript{20} The Response-Window Priming Task (Draine & Greenwald 1998) seeks to measure the strength of association between two constructs by asking participants to classify, as accurately as possible, certain stimuli within one of two categories. For example, in the case of Van Bavel & Cunningham (2009), the participants were asked to classify words as belonging to the ‘Good/Liked’ or ‘Bad/Disliked’ category, for which they had a 525ms response window. The key element here was that a photo of a person (who was a member of their own team, a member of the opponent team, or a non-member) was flashed for 150ms just before each word appeared. This allowed the researchers to measure the effects of each photo on performance accuracy. The gathered data showed that faces with positive (in-group) associations increased the accuracy of positive-word categorization and decreased accuracy of negative-word categorization.
visible. Then (by means of a Lexical Decision Task, which measures the speed of recognition of stereotypic vs. counter-stereotypic words) researchers found that gender stereotypes were activated only in the group whose task was to distinguish animate from inanimate objects, and not in the others.

In another study (Mitchell et al. 2003, experiments 1–3), researchers found that an ambiguous stimulus (a liked Black athlete or a disliked White politician) activated either positive or negative evaluations depending on whether the task required participants to direct their attention toward profession or race features. Further (experiments 4–5), after finding that White women tend to associate the ‘White’ and ‘female’ categories with positive evaluations, and the ‘Black’ and ‘male’ categories with negative evaluations, it was also shown that the same ambiguous stimulus (a White man or a Black woman) can elicit either positive or negative evaluations depending on whether the task that participants engage with makes either race or gender salient.

It would be wrong to conclude from studies like these that automaticity is always dependent on attention—it is not. But it would also be erroneous to hold that automaticity is blind to, and independent from, attentional activity. As Blair puts it, “[t]he fact that a process does not need the perceiver’s attention to operate does not necessitate the conclusion that attention cannot influence that process” (2002, 252).

What can be concluded from this? It seems that the conditionality of automaticity is so broad in scope that conceiving automatic processes as mechanical is a misdescription. If reflexes are indeed triggered by the mere presence of the relevant stimulus, then this is a way in which automatic processes are (at least sometimes) unlike reflexes, since in many cases there is nothing like a clearly identifiable stimulus whose mere presence almost invariably activates the same automatic process. The latter, on the contrary, are sensitive not only to a specific triggering condition, but also to the agent’s motivational states and occurrent goals.

### 3.3.2. Are automatic processes ballistic?

The ordinary sense of the word ‘automatic’ has several meanings, and this polysemy is not absent from the theoretical usage: a cognitive process can be called ‘automatic’ if it is involuntary, effortless, unconscious, unintentional, triggered without the participation of working memory, hard to correct or alter… Despite its homonymy, several researchers have proposed that automaticity’s focal meaning should be its **ballistic** character, i.e. its being unstoppable once initiated. Regardless of whether a process is unconscious, unintentional, effortless, etc., if the process cannot be stopped or deflected once started it is automatic. This criterion seems to make
sense from an experiential perspective: when we see a familiar word we cannot help but read it, and when we see a person’s face we cannot but recognize it as such.  

The psychological construction of automaticity as ballistic has had its impact on philosophy. Strawson (2003), for instance, has argued that most mental activity is ballistic: every case of emergence of mental content (as in perceiving, imagining, remembering, judging, drawing a conclusion, making a decision, forming an intention…) relies on automatic processes, and is therefore produced by processes that once started run to completion without the smallest possibility for us to interfere with or stop them. Wu (2013a) considers Strawson’s ‘mental ballistics’ to be just another name for ‘automaticity’, and takes his argument even further by holding that if all mental content-production is ballistic, then even things Strawson considers non-ballistic and intentionally produced (e.g. stopping a chain of thought or redirecting attention) cannot escape from automaticity, so that the threat that mental ballistics poses for action control extends throughout the whole spectrum of mental activity. And this is a threat to control, for a ballistic process is by definition inalterable and incorrigible. We cannot change a rock’s trajectory once we have let it go from our hand, to use Aristotle’s example.

Further, Gendler (2008a, 2008b) has suggested that one feature of automatic-associative mental states is that they are recalcitrant to available evidence, so that their behavioural output cannot be cancelled even when we have clear evidence for its inappropriateness. She mentions examples like the fear that even wise persons feel when hanging from a precipice despite being certain that the cage they are suspended in is completely safe, and the unavoidable disgust caused by the prospect of eating delicious, yet feces-shaped, chocolate fudge. Automatic processes are thus held to be ballistic also in the sense of being insensitive to relevant, available evidence. And this epistemic insensitivity is yet another reason why automatic processes seem to escape control, since they cannot be corrected mid-performance on the basis of recently updated information.

---

21 For Shiffrin & Schneider, “[s]ome automatic processes may be initiated under subject control, but once initiated all automatic processes run to completion automatically” (1977, 160). Norman and Shallice (1986) consider that, since automatic processes run to completion once triggered, any kind of error-correction requires deliberate attentional control. Bargh refers to this as automaticity’s “autonomy” (1992, 186), and suggests that it is its one core feature. For Stanovich, “TASS [i.e. intuitive] processes cannot be turned off or interfered with by central systems. Their operation is obligatory when triggered by the relevant stimuli […]. TASS processes tend to be ballistic—they run to completion once triggered and cannot be aborted in midcourse” (2004, 39). (See also Logan & Cowan 1984; Moors & De Houwer 2006, 301–302.)

22 “[I]f we consider things plainly, we find, I think, that most of our thoughts—or thought contents—just happen. […] Contents occur, spring up—the process is largely automatic. Even when our thoughts are most appropriate to our situation and our needs as agents, action and intention need have little or nothing to do with their occurrence” (Strawson 2003, 234).
The ballistic interpretation of automaticity has led dual-process theorists to resort to a top-down, intellectualist account of control: the only way to intentionally direct automatic processes is by either reflectively overriding their influence when they are misguided, and then replacing them with more appropriate reflective processes; or by reflectively shaping one’s habits diachronically, in order to reshape the unwanted automatic associations towards patterns that better reflect our higher-order goals, beliefs, and commitments (Stanovich 2004; Gendler 2008b; Fridland 2015). The ballistic interpretation of automaticity thus leads to the intellectualist view that control over automatic-associative processes require top-down reflective guidance and structuring—a view quite similar to Aristotle’s claim that care for one’s habits is a necessary condition for agency and control (as argued in Chapter 1).

But do we have to always reflect in order to readjust an automatic process? Think about it. Experiential reflection seems to suggest, on the contrary, that automatic processes are often automatically self-correcting. When you over-squeeze a plastic cup, you immediately readjust the strength applied by your hand, with no need for reflection to get involved. As you walk, run, skate, cycle, drive, or swim, you automatically adjust your posture on the basis of visual and vestibular cues, in ways so complex that explicit reflection would be unable efficiently to specify, and that rely instead in associations acquired through repeated experience. In fact, automatic error-correction is so fast and efficient that sometimes we are completely unaware of it.

For instance, an experimental subject, sitting in the dark, stares at a single light source and points a finger at it. Suddenly a new light appears in the periphery and the observer has to move his/her eyes and finger rapidly towards this target. While the eyes are in transit, the light is moved a little bit to the left or right. Neither eyes nor finger have a problem in correcting for this and both end up right on target. Yet the subject does not see the extra target motion, even though the oculomotor system automatically corrects for it. All that is perceived is a flashing target and the eyes and hands making a movement towards it. (Koch & Crick 2001, 893)

Thus, not all error correction processes must call reflection for help: automaticity has a norm-sensitivity, an ability to adapt to environmental changes, a capacity of its own to detect and resolve tensions present within the dynamic stream of associative activity (Rietveld 2008; Brownstein & Madva 2012; cf. §2.2 above). You probably just performed an automatic error correction in reading this paragraph, by reading ‘environmental’ as ‘environmental’ in the second line.
Case 3: Automatic error correction

Rabbitt (2002) compared people’s abilities to correct, report, and recall errors in a simple task. Participants (both young and old adults) were instructed to look at a screen split into four squares; when a dot appeared in each of the squares they had to press a corresponding button. One group of participants was instructed to immediately correct the mistakes they made during the task; a second group was instructed to press a fifth button every time they recognized a mistake; a third group was randomly interrupted and asked whether they remembered having made a mistake in the last three responses; lastly, a fourth group was told to ignore all errors and just keep responding. The length during which the stimulus appeared (i.e. the Response Signal Interval, or RSI) varied randomly throughout the task between 150ms, 500ms, 800ms, and one second.

The first relevant finding was that for all participants the response that followed a mistake was slower than the previous ones. This slowing occurred even when they were unable to report or recall the errors. Further, participants in the first group were remarkably fast and accurate in correcting their errors: throughout all RSI’s, errors were on average faster than correct responses, and error corrections were on average even faster than errors.

Additionally, participants displayed much more accuracy in error correction than in error report or error recall. Across all conditions, it took participants much longer to report an error than to correct it, and failed to report errors much more often than they failed to correct them. Given that error correction can occur notably quickly (as quickly as 40ms after the error is performed [Rabbitt 1966a, 1966b]), Rabbitt concludes that error correction cannot depend on reflective recognition and rectification processes. The error-correction mechanism must therefore be automatic. Crucially, these automatic error corrections always worked to increase response accuracy, and never turned a right response into a wrong one (2002, 1082).

So automatic processes can be sensitive to error, and are able to perform quick and efficient self-correction without the intervention of slower and coarser-grained reflective processes. This reveals that Type 1 processes are not (or not necessarily) ballistic phenomena, insensitive to available evidence. Case 3 also illustrates how automatic processes can be norm-sensitive, since they can reliably guide behaviour away from error towards correction and accuracy. I will say more about the nature of this norm-sensitivity below.

It thus seems too harsh on automatic cognition to call it “unable to keep pace with variation in the world or with norm-world discrepancies” (Gendler 2008b, 570), even if it does behave like this in some cases, or with respect to certain kinds of evidence—particularly when it is in the form of novel propositional knowledge. We should rather agree here with Brownstein and Madva in claiming that automatic processes “can be norm-sensitive in virtue of their

---

23 For other cases of automatic error detection and correction in sensorimotor processes, see Gordon & Soechting (1995), Rabbitt (1990), and Logan & Crump (2010). For automatic error detection in cognitive tasks, see Fernandez Cruz et al. (2016).
responsiveness to affective states of disequilibrium. Responsiveness to such affective states is flexible, self-modifying, and [...] a genuinely normative phenomenon” (2012, 428). Those affective states of disequilibrium can, of course, be misguided with respect to the overall available evidence (as in the case of the trembling wise man hanging over a precipice from a perfectly safe cage), but this is not whole story about automaticity, since such states of disequilibrium, far from being ballistic and unalterable processes, are also a part of a reliable dynamic system of on-line error detection and correction. This should raise new doubts concerning the interpretation of automatic processes as blind and lacking control.

3.3.3. Are automatic processes unconscious?

To a large extent, what makes psychological findings on the pervasiveness of automaticity worrisome for a theory of action is that the workings of automatic processes often seem to be inaccessible to consciousness. For if just about any bit of our environment can alter our behaviour behind our backs, the control over our lives that we otherwise feel we have seems merely illusory. This is the kind of concern that makes automaticity unbearable, to use Bargh and Chartrand’s (1999) expression. So, are Type 1 processes unconscious?

Some distinctions

Since the words ‘conscious’ and ‘unconscious’ are extremely ambiguous, some conceptual clarification is in order. First, to we should distinguish the question whether automatic processes are access conscious (i.e. whether they are accessible for reasoning, action guidance and verbal report) from the question whether they are phenomenally conscious (i.e. whether there is something it is like to have such processes) (Block 1995). We know we have phenomenal consciousness of automatic processes insofar as some of them express themselves in the form of intuitions, hunches, insights, emotions, motivations, pulls, or drives (Evans 2010a). But we know as well that such phenomenal consciousness extends only to the process’ outcome: the experience of intuitions itself, for instance, implies that its originating process is unclear.

24 Classical conceptions of automaticity link it to absence of awareness, and reflective processes with consciousness. Shiffrin & Schneider consider that “[t]he phenomenological feeling of consciousness may lie [...] particularly in the subset [of short-term memory] that is attended to and given controlled processing” (1977, 157); they thus relate awareness to processes of control, as opposed to automaticity, although they do not consider all control processes to be conscious. For Posner & Snyder (1975), one of the trade marks of an automatic process is that it occurs “without giving rise to any conscious awareness”. Lack of awareness is labelled one of automaticity’s “four horsemen” (Bargh 1994). In philosophy, Suhler & Churchland (2009) criticize the idea that all control is conscious, and use the terms ‘automatic’ and ‘unconscious’ interchangeably. Fridland (2015) critically examines the automatic-unconscious link, pursuing a similar line than the one I develop here.
Moreover, having phenomenal consciousness of a process or state is largely tangential to the question of control: people with vertigo, for instance, are phenomenally aware of their processes, but this experience does not bring them closer to controlling their vertigo. The same can be said of the temptations that haunt akratic agents or the slips of action that undermine our concurrent plan. Access consciousness, on the other hand, seems to be relevant for the question of control, insofar as our control over a mental state or process depends on our ability to inscribe it within a space of reasons, in which it can be rationally evaluated and introduced in processes of deliberation or planning. And being access-conscious of a process or state means we are able to report it and bring it into our reflection. Thus the crucial question here is whether automatic processes are access-conscious so that they can enter into reflective processes.\footnote{There is a subsequent ambiguity, since one can be access-conscious of a process’ origins, of its content, or of its impacts on further processing. This issue will be discussed below.}

Second, and in relation to the last remark, the claim ‘We lack access to our automatic processes’ may mean two different things. It may mean that automatic processes can, or even tend to, work without our being access-conscious of them; or it may mean that they are altogether \textit{inaccessible} to consciousness. I call the former the \textit{limited-access claim}; and the latter the \textit{inaccessibility claim}. The limited-access claim seems right, because access consciousness is a limited resource (often linked to the limits of attention and working memory capacity), and automatic processes are so abundant that most go by unnoticed. But you can endorse the limited-access claim and still be hopeful about the possibility of conscious control of automatic processes, because we can at least in principle direct our limited introspective resources to the associations that matter most to us in some normative or practical sense, gain access-consciousness of them, and include them in our processes of reflective control. But the much stronger inaccessibility claim, if true, would take such hopes away.

\textit{Inaccessibility: For and against}

Gendler (2014) has recently defended the view that some automatic processes are inaccessible, and nevertheless exert a powerful influence on our behaviour.\footnote{She argues that they have no “distinctive phenomenology”. Although this is a claim about phenomenal consciousness, I take it that her worry is also about the lack of introspective access.} In her view, this holds for the automatic activation of associations between social categories (like gender, race, and social status) and evaluative categories that have been built into our intuitive minds through repeated experience of the dominant norms and values of our culture, and that now tend to modulate our perceptions, evaluations, and behaviours, despite any reflective opposition to them. Taking
Plato’s image of the divided soul as composed of a rational charioteer and the two irrational
horses he tries to control, Gendler considers that the relationship between the charioteer and his
horses is analogous to that between reflection and intuition; but she argues that the image is
incomplete, because there is yet another horse, namely the pull of automatic processes that, like
implicit associations, have no phenomenology and are inaccessible to introspection. The hidden
power of such third-horse associations reveals how limited reflective knowledge is as a master of
our behaviour: it can attempt to control whatever we have some epistemic access to; but
whatever we do under the influence of implicit associations falls outside of the scope of control,
since we have no awareness of their influence.

So Gendler defends the inaccessibility view on the grounds that implicit associations are
inaccessible. Now, are they inaccessible? The main evidence for this is the discrepancy, often
experimentally revealed, between implicit measures of attitudes and self-reports. Explicit self-
reports rely on introspection, but implicit measurements (like sequential priming or Implicit
Association Tests) seek to detect the presence of associations between categories while
bypassing introspection (Gawronski et al. 2007, 181). Since they are introspection-independent,
implicit measures have been assumed to be ‘gateways to the unconscious’ (Hall & Payne 2010),
particularly when a mismatch is found between self-reports (i.e. what the person explicitly and
willingly admits she thinks) and implicit measure results (the person’s mental associations that
are experimentally found to be activated).27 The findings that, for instance, people’s implicit
attitudes toward women or racial minority groups show low correlation with their self-reports,
and that implicit attitudes can significantly shape their decisions (see Case 4), is taken by
Gendler as evidence that invisible, implicit associations (e.g. the bias towards considering men
more competent than equally qualified women) can be activated without the agent’s awareness,
and that reflective knowledge (e.g. that people’s gender is irrelevant to judgments about
competence) is powerless to counter their invisible influences on behaviour.

Case 4: Implicit bias in academia
The introduction mentioned the work by Moss-Racusin et al. (2012) that demonstrated professors
evaluated a male candidate as more competent than a female candidate, the only difference between
them being their name. In relation to this, but specifically in the field of philosophy, Healy (2013a;
2013b) investigated which works and authors received the most citations in some of the most

27 Mismatches between self-reports and implicit-measures of attitudes are very common: in a meta-analysis,
Hofmann et al. (2005) report a mean implicit-explicit attitude correlation of only .24. Bear in mind that in this
context ‘attitudes’ does not have the philosophical sense of ‘propositional attitudes’, but rather the social-
psychological sense of a stable set of associations between a certain object and a valuation (cf. e.g. Cunningham &
Zelazo 2007).
prestigious anglophone philosophy journals (Mind, Nous, the Journal of Philosophy, and The
Philosophical Review) from 1993 to mid-2013. He found that of the 200 most cited works in the
period, nineteen were written by women. This amounts to 3.6% of the total. The works of a single
man almost double that number (references to David Lewis account for 6.3%). After ruling out other
explanations, Healy claims that this is due not to women’s underrepresentation in the field, but rather
to the fact that “women publish, yet their work is not cited”. Implicit bias has been held as the cause
of such patterns (Saul 2011). Implicit bias is, of course, not unique to academia. A review gathers
analogous documented cases with respect to race, ethnicity, nationality, gender, social status, etc.,
tacitly affecting the decisions and actions of nurses, doctors, police officers, employment recruiters,
and several other groups (Jost et al. 2009).

As mentioned before, Gendler’s interpretation of these studies, and her claim about the existence
of an invisible horse, relies on the claim that implicit attitudes are inaccessible, and the crucial
evidence for such claim is the discrepancy often found between implicit attitudes and self-reports.

Nagel (2014) has argued against the latter by referring to the findings (reviewed in
Gawronski et al. 2006) that discrepancies between implicit measures and explicit reports can be
eliminated. Such discrepancies seem to be largely due to the fact that self-reports are influenced
by motivations (like the intention not to be perceived as a prejudiced person) and opportunities
for deliberation (i.e. allowing participants time to reflect upon their evaluation, instead of
reporting their immediate ‘gut feeling’ about the subject). In situations in which said factors are
controlled (i.e. when participants have no motivation to hide prejudices, and their report is fast
and spontaneous), implicit and explicit attitudes are much more closely aligned. This suggests
that people actually can have conscious access to their automatic associations, attitudes and
evaluations. It would thus seem appropriate to agree with Nagel that “[g]ut feelings, even if not
endorsed, are available to introspection” (2014, 234; see also Levy 2014). Nagel derives further
support from the fact that the subjects of the Moss-Racusin study who tended to give the lowest
assessments of the woman’s competence also tended to give the most sexist responses to
questions in the ‘Modern Sexism Scale’ (Nagel 2014, 231–232). Nagel argues that, since those
with the most inaccurate and biased beliefs about the situation of women tend to produce the
most biased evaluations in the Moss-Racusin study, a person’s explicit and implicit attitudes do

28 In an earlier study, Haslanger (2008) reviewed a broader sample of prestigious anglophone philosophy journals,
and found that in the period between 2002 and 2007 12.36% of the total published articles was authored by women.
29 The Modern Sexism Scale, in the version prepared by Swim et al. (1995), assesses sexist attitudes by asking
people whether they agree with statements like “On average, people in our society treat husbands and wives equally”
and “Over the past few years, the government and news media have been showing more concern about the
treatment of women than is warranted by women’s actual experiences”.
seem to line up, and therefore that there is no evidence that there are inaccessible automatic processes, and no need to posit an invisible third horse.

Nagel argues that such introspective availability is sufficient to dispel Gendler’s worry about the reflective uncontrollability of implicit attitudes. According to her view (much closer to the Socrates of the *Protagoras* than to the Socrates of the *Phaedrus*), knowledge *does* have a commanding power over automatic associations, since the introspective accessibility of implicit attitudes gives “a suitably pensive person” the chance to question herself about the discrepancy between her gut feelings and her reflective commitments. In other words, the fact that implicit attitudes are consciously accessible makes them reflectively controllable, since they can enter the space of reasons and deliberation.

This much certainly seems to be the case: if you have a sufficiently reflective disposition, you can question yourself about such intuition-reflection inconsistencies, and decide to act in response to them. However, does this imply that your knowledge of your inconsistency, and your subsequent self-examination, empower you to control the influence that biased automatic associations have over your behaviour?

**Impact awareness—or the lack thereof**

In order to answer this question, it is useful to make yet another couple of distinctions between different senses of the word ‘consciousness’ (which I am taking to be synonymous with ‘awareness’). Following Gawronski et al. (2006), there can be said to be three ways in which we may have, or lack, awareness: (1) *source awareness*, i.e. awareness of the implicit process’ causal origins, (2) *content awareness*, i.e. awareness of the specific associations that constitute the automatic process; and (3) *impact awareness*, i.e. awareness of how or when the implicit process influences latter psychological processes. It is important to draw this distinction, since explicit and implicit attitudes may differ in conscious accessibility in some of these senses, even if they do not in the others. In fact, after a review of the literature Gawronski et al. found that, whereas there is no difference between implicit and explicit attitudes with respect to their source awareness (we are often unaware of the origins of our attitudes, regardless of whether they are implicit or explicit) and content awareness (the contents of our attitudes are introspectively accessible), nevertheless there is a significant difference with respect to impact awareness. Evidence suggests that we lack awareness of the effects that implicit attitudes have on subsequent psychological processes.
The classical case for impact inaccessibility comes from Nisbett and Wilson’s (1977) work on introspection (see Introduction above), which suggests that, although the environmental features that lead to our decisions are certainly accessible, the impact that they have on our decision may not be so. In Nisbett and Wilson’s stockings case, introspection seems to be dedicated to confabulating plausible explanations for our behaviour, rather than correctly identifying its causes; and many other studies suggest the same (Wilson 2002; Carruthers 2009, 2011). Evans considers that phenomena like belief bias (see Introduction above) also support the lack of access to intuitive processes’ impact, since in these cases we may be very motivated to reason logically, both to comply with the instructions given and to feel good about how smart we are. In spite of all this, our intuitive minds often exert control and hijack our reasoning. Certainly, no participant in a reasoning experiment reports that belief bias is the basis for their answer! (Evans 2010b, 117)

Further, Gawronski and his colleagues offer an interesting argument in support of the claim that we lack impact awareness of implicit attitudes (see Case 5).

**Case 5: The inaccessibility of implicit attitudes’ impacts**

How can we determine whether an association’s impact is accessible? According to Gawronski et al. (2006), if a person is (i) aware of the potential impacts of a biased implicit association, (ii) sufficiently motivated to act in non-biased ways, and (iii) cognitively capable of overriding such influence, then she will override it. Given this, if people who are sufficiently motivated to override such influence, and are cognitively able to override it, nevertheless fail to do so, then it can be concluded that they lacked impact awareness.

In a study conducted in Germany, researchers measured the extent to which participants’ implicit bias against Turkish people affected their categorization of ambiguous behaviour (Gawronski et al. 2003). Participants had to read a story in which a person (called “A”) engaged in morally ambiguous behaviour; and then had to assess how negative A’s behaviour was (rating it from 1 to 5 according to negative terms like “objectionable”, “insensitive”, and “arrogant”, and also predicting how likely A was to engage in unambiguously negative behaviour). One group of participants was led to believe that A was German, while the other was led to believe he was Turkish. Unsurprisingly, people’s implicit attitudes toward the ethnic groups influenced how negatively the ambiguous behaviour was evaluated: participants with strongly negative associations toward Turkish people identified the behaviour as more negative when A was portrayed as Turkish than when portrayed as German, whereas for participants with more neutral associations A’s ethnic origin did not affect how negative his behaviour was taken to be. (Implicit associations were measured via IAT.)

Importantly for current purposes, the study also measured participants’ motivation to control prejudiced behaviour, and found that high motivation to control prejudice did not override the impact
of negative associations on the moral evaluation of A’s behaviour, since participants with very negative implicit associations toward Turkish people gave significantly more negative evaluations of A when he was portrayed as Turkish than when he was portrayed as German, regardless of whether they were motivated to control prejudice or not. What participants’ motivation to control prejudice did modulate was their explicit reports: those with low motivation to control prejudice provided explicit reports very much in line with their levels of implicit prejudice, but those with high motivation to control prejudice gave explicit reports low in prejudice, even when their implicit associations revealed prejudice (as one would expect, given the aforementioned influence of motivation on explicit reports). In short, motivation to control for prejudice led people to confabulation about their attitudes, but it did not lead them to stop their implicit prejudice from modulating their behaviour.

The fact that people less motivated to control prejudice were able to give explicit reports in line with their implicit associations suggests that the content of such associations was introspectively accessible. Now, since in this experiment participants were neither under time pressure nor asked to perform other cognitive-resource-demanding tasks, it can be assumed that those motivated to control prejudice had their working-memory resources sufficient free to perform an override of an automatic process’ influence on subsequent behaviour (in this case, the moral evaluation of A’s behaviour), had they been aware of such influence. The fact that highly motivated people with available working memory capacity did not override said impact strongly suggests that the potential impacts of their implicit associations were not introspectively accessible to them, even though the content of such associations was. Thus, the impact of implicit attitudes on subsequent mental processes seems to be inaccessible.30

There are several potential objections to this argument. One of them is that perhaps participants did not actually have the power to override the biased association’s impact. If this was so, it would not follow that impact was inaccessible to them. But we have no reason to suppose that participants lacked the executive control required to inhibit predominant automatic processes. For, as mentioned above, working memory (understood in the broad sense [§2.2]) enables the ability to inhibit predominant automatic processes inconsistent with current goals, and allows for agents to exercise self-control in the face of temptations.31 In this particular case, there was no identifiable element of the situation that could be said to have depleted working memory capacity: they were in

30 Hall & Payne (2010, 229–235) discuss a number of other studies suggesting the introspective inaccessibility of associations’ impact.

31 If the narrow sense of ‘working memory’ is assumed, then it is not the executive function that is relevant here. The one that is is inhibitory control, whose task is to stop features irrelevant or contrary to the task from capturing attention or influencing behaviour (Miyake et al. 2000, Diamond 2013). We employ inhibitory control e.g. to keep loud noises or other people’s conversations from distracting us from our current task, or to stop ourselves from blurtting out the first thing that comes to mind in conversation when we notice it may be inappropriate. Like the other executive functions, inhibitory control has limited capacity and can be overwhelmed if we need to inhibit too many things at once. As mentioned in §2.2, the broad conception includes inhibitory control as one of working memory’s attention-relative functions.
a calm environment, not pressed for time, and not previously or simultaneously burdened with any other cognitively demanding tasks. So their executive working-memory resources were available to attend to and override any predominant responses that were deemed inappropriate—like acting in a prejudiced manner due to an implicit association. But they did not override their implicit association’s impact on their evaluations of A. This leads to the conclusion that they did not have access to the association’s impact—they simply couldn’t see what was going on.

It may be objected, however, that they were unable to recognize the association’s potential impact not because such impact is usually inaccessible, but rather because of their motivation to control prejudice, since such motivation was shown to lead people to confabulate rather than accept the existence of a biased association in the first place. Paradoxically, then, people low in motivation to control prejudice would have a better shot at overriding prejudice than those highly motivated to control it. But the impacts of implicit attitudes would not necessarily be inaccessible. I think nothing said so far proves this position wrong, but the burden of proof should be on the objector, given the substantial body of evidence for the limits of introspection with respect to identifying the actual causes of our behaviour, and our tendency to engage in confabulation (§1.1.1.; Wilson 2002; Carruthers 2011). At any rate, what the evidence from Case 6 seems to show is that we are more susceptible to confabulate reasons for our behaviour when we are most interested in catching the impacts of our automatic associations, because it is in those situations that our explicit attitudes are more prone to diverging from our implicit attitudes. This higher tendency for reflection to engage in confabulation suggests a strong limitation in its ability to control the effects of discordant implicit associations.

If all this is correct, then although we have conscious access to the content of automatic-associative processes (often in the form of intuitions, gut feelings, and emotional reactions), we nevertheless seem to lack conscious access to their impacts on subsequent cognition and behaviour. It follows from this that even people who are conscious of having an implicit bias toward a certain social group, and who are reflective enough to notice the conflict between their implicit biases and their reflective commitments, may nevertheless fail to cancel the impacts of their implicit bias due to impact inaccessibility. A person may, for instance, know that she is biased against old people, despite reflectively rejecting such bias and being reflectively motivated not to act on the basis of it, and still fail to control her bias’ impacts on “micro-behaviours”—things like the distance she keeps from old people in conversation, the tone of her voice, her facial expressions, or the amount of eye-contact she makes—simply because bias invisibly modulates these micro-behaviours. Situations like this are what Gendler (2014, 206–207) is concerned about, and in relation to them content awareness does not seem to allow us to reflectively control our own behaviour—or at least not
synchronously, i.e. by reflecting on the spot. There may be ways in which we can use reflection to
diachronically control the invisible influence of bias—by, e.g. re-shaping our habitual associations,
or rearranging our environment so that bias has fewer chances of effectively shaping decisions and
actions. I will say more on extending control diachronically later (see Chapter 5). For now the
crucial point is that our lack of impact awareness renews Gendler’s concern over the limits of
reflection’s power to control automatic processes. This is because successfully controlling a biasing
influence depends not only on being able to recognize its presence, but also, and crucially, on our
ability to identify how it may exert its influence.

Blair et al. (2004) report a sobering illustration of how content awareness seems severely
limited as a tool for reflective control in the absence of impact awareness. By randomly sampling
inmates in United States prisons, Blair and her colleagues determined that Black and White
inmates with equivalent criminal records received roughly equivalent sentences. This sounds like
good news, except that in the same sample it was found that inmates with more “Afrocentric” (sic)
facial features (e.g. broader nose, thicker lips, curlier hair) received significantly harsher sentences
than those with equivalent records but less “Afrocentric” facial features, regardless of whether they
were Black or White. The problem with having content awareness and no impact awareness is that
you know which bias to look for, but may not know where to look for it. In fact, it may even be
impossible to identify all the different instances or forms that implicit bias takes.
§3.4. Automatic intentionality

In the previous section I argued that the labels ‘mechanical’, ‘ballistic’, and ‘unconscious’ do not apply to automatic-associative processes. Against the view that automaticity is a mechanical, stimulus-determined process, I presented evidence suggesting that the activation of automatic processes is conditional on the agent’s broad practical situation (i.e. responsive to multiple environmental features and the agent’s motivational states and direction of attention), and conditionality extends to such an extent that the idea that a clearly specifiable input almost invariably activates a given association cannot capture automaticity’s fine-grained sensitivity to the situation’s demands.

Against the ballistic view, I provided evidence suggesting that automatic processes display error-detection and self-correction properties so quick and efficient that they could not be attributed to reflective control. These reflection-independent error-detection and -correction capacities reveal that automaticity has a norm sensitivity of its own—not based on slow, serial reflective processes, but rather grounded on a pre-reflective responsiveness to environmental tensions, a sort of habitual sensitivity closely related to anti-intellectualist concepts like Merleau-Ponty’s motor intentionality, Dreyfus’ skilful coping, Clark’s pattern-completing processes, Rietveld’s directed discontent, and Brownstein and Madva’s normativity of automaticity. In fact, in the notions of ‘intuition’ and ‘automaticity’, precisely characterized through dual-process theories, provide a contemporary and updated incarnation of the notion of operative intentionality that inspires and informs all versions of anti-intellectualism. From this perspective, automaticity is a kind of intentional engagement with the world that is prior to the reflective, top-down, intention-coordinated interaction proper to operative intentionality (§2.1.2A–B).

To further clarify the nature of this automatic intentionality, consider the difference between an explicit, episodic memory about a certain event, and the kind of implicit memory that constitutes habits and skills (Evans 2010b, 62). An episodic memory is susceptible of verbal report and accessible by reflective processes, but a habit or a skill is not (you can say you know how to play the piano, but the propositions you use to describe your skill do not constitute that skill).32 Further, whereas explicit memories register ‘variances’, i.e. the features that are distinctive of the remembered situations, automatic and implicit memories (i.e. associations, habits, and skills) encode ‘invariances’, i.e. the underlying form or pattern that constitutes a type of behaviour (Sherry & Schacter 1987). In learning to talk, children do not learn the specific sentences they hear; they abstract the underlying forms. In learning to play a scale in the piano, we obviate the particularities of each time we play the scale and retain the abstract, normal

32 Stanley & Krakauer (2013) oppose the latter claim. For further discussion on the nature of skill see Chapter 4.
sequence of finger and hand gestures. In social categorizing, we automatically assess which of
the stereotypical categories the person in front of us can fall into. Automatic processes therefore
work on the basis of the invariant features that have been extracted from repeated experience of a
given practical environment and stored into valenced associations. This is what allows
associations, habits, and skills to be transposable to an infinity of particular situations (like
Merleau-Ponty’s organist and his ability to play a different organ with very little practice time
(§2.1.2B)). Thus, associations encode forms or patterns of behaviour. This formal aspect of
automaticity is what makes it a normative phenomenon, since automatic-associative processes
can detect deviations from the form and dynamically guide the behaviour toward the pattern
immanent in the active associations. Insofar as the results obtained in this chapter are
generalizable to automatic processes, it can be said that the normativity of automaticity consists
in the pull toward a norm that has been stored in the associative networks as a certain (cognitive,
social, affective, evaluative, or sensorimotor) form.

The existence of this intentional and normative aspect of Type 1 processes challenges one
of the classical tenets of modern psychology, namely the opposition we started this chapter with
between automaticity and control. Automatic processes, after all, have also turned out to
embody a certain kind of control that is independent from reflection. This turns out to challenge
the intellectualist idea that control is a necessarily hierarchical phenomenon that requires for
lower-order cognitive or sensorimotor processes to be shaped in accordance with higher-order
goals, intentions, or reflective processes like deliberation or planning. Automaticity’s operative
intentionality opens the possibility for a kind of control that is not hierarchical, not determined
by correspondence to higher-order intentions or goals, but rather constituted by an agent’s on-
line skilful interactions with her practical environment, guided not towards the completion of a
goal, but towards the alleviation of a tension, and the completion of an associative form.

Dual-process studies of the different forms of automaticity and their influences on
cognitive and social behaviour (cf. Introduction above) show that, despite being normative and
intentional, the automatic alleviation of tensions and completion of forms can often lead us to
errors and biases. Acting on the sheer basis of automatic associations is tantamount to acting
along the patterns established by experience. Insofar as this experience is biased by the
limitations proper to the human mind’s evolutionary history or one’s own cultural baggage,
automatic intentionality is susceptible to both cognitive and practical errors, and could lead to
stereotypical reactions rather than sensible actions (Montero 2015).

A question stems from this description of the intentionality of automaticity: is automatic
control sufficient for the production of full-blown intentional actions? Chapter 4 addresses this
by focusing on the kind of behaviour that seems most representative of automatic control (and least explainable by reflective control): skilled action.

Finally, regarding the relationship between automaticity and consciousness, I have provided evidence suggesting that we sometimes have no access to the impacts of automatic processes on further cognitive processes. I have argued that this limit of our conscious access to automatic processes has potentially worrisome normative consequences, because the lack of impact accessibility implies a limitation in reflection’s power to synchronically control automatic processes, and suggests the need to extend control over the effects of associations diachronically. This is the main concern of Chapter 5.
Chapter 4: Automaticity and control in skilled action

Niki Nakayama is a world-renowned chef. The food she makes in the L.A. restaurant n/naka has the paradoxical qualities of strict observance of the Japanese kaiseki tradition and willingness to stylishly break the rules. Commenting on the way she works, Nakayama says:

Cooking is the one thing that I feel I can completely trust what I'm doing. When I'm planning a dish, my mind is completely shut off. It's all based on feeling: 'this has to be here, this has to be here, this feels right here, this looks right here.' I think it's similar to that meditative state that people can get to, where they're not listening to their minds anymore, but it's just that moment.

Her description is remarkably alike M. Csikszentmihalyi's description of flow, the kind of action proper to highly skilled performers at the top of their form: “in flow—he says—there is no need to reflect, because the action carries us forward as if by magic.” (1990, 155). But this reflection-less property of flow contrasts with another of Nakayama's comments:

The best advice that I've been given was to never stop learning. Because the moment you give up and think you know everything, you're done as a chef. You should just quit.¹

Nakayama's account of what she does elicits a theoretical puzzle. What use is it for an expert to keep learning, if while acting she is not supposed to listen to her mind? For, if skilled action is guided merely by what intuitively feels right in the moment, then the coarse-grained reflection required for learning something new could hinder performance. These are, then, crucial questions about the nature of skill: (1) Is automatic control sufficient for expert action, or must it include an element of reflection? And (2) if the latter, then how can reflection be integrated with the over-practiced, automatically-controlled routines?

As to the first question, anti-intellectualists take the first horn of the dilemma: the more skilled you are, the less you need to think about what you do or how to do it: the practical situation simply solicits the proper response from you (§2.1.2 D–E). As in the Daoist notion of wei-wu-wei,² the expert 'acts without doing': responds without planning or calculating, by

---

¹ Nakayama’s remarks are taken from documentary series Chef’s Table (chapter 4, dir. D. Gelb).
² See e.g. Loy (1985); Slingerland (2003).
merely resonating with the dynamics of her practical environment, with which she forms a harmonious system. In contrast, Intellectualists go for the second horn of the dilemma, often pointing to the constant drive for improvement as evidence that reflection is a crucial aspect of expertise. This drive, the intellectualist argues, is precisely what tells experts apart from other competent performers: their continuous search for ways to improve, to enrich their performance, leads them to refine their activity in ways that would be impossible if it was fully automated. Complete automaticity begets stagnation, so experts must keep a window, however small, for reflection to continue assessing and reshaping automated routines, thus improving control over performance. According to several intellectualists, this reflective drive for improvement is what separates gold from bronze medalists, top performers from very good ones.3

In short, anti-intellectualists claim that reflective control becomes unnecessary (and ultimately a hindrance) as an agent acquires greater levels of skill; but intellectualists contend that reflective control is necessary for skilled performance, no matter the skill level, and becomes crucial in reaching the highest levels of expertise. Nakayama’s claim that while cooking her action is based not on thinking but solely on feeling supports the anti-intellectualist’s view; but her further statement that a chef must never stop learning supports the intellectualist’s position. So what kind of control do skilled and expert agents employ in performing skilful actions?

In this Chapter, I will (§4.1) review the recent philosophical debate around the topic of skilled action. I will then present a contribution to the debate by examining three psychological phenomena that anti-intellectualists have recently argued are unexplainable through intellectualism: (§4.2) choking under pressure, (§4.3) expertise-induced amnesia, and (§4.4) experts’ tendency to confabulate. Do these phenomena really give support to the anti-intellectualist approach?

Intellectualists have a tough task ahead, because skill seems to have been the most recalcitrant kind of action for intellectualist explanation. And even if they manage to put forward a case for the view that reflection is necessary for skilled action, they must also answer the second question posited above: how is this reflective element integrated with the automatic processes that control the fine-grained cognitive and motor routines of skilled agents?

---

3 See e.g. Annas 2011, 17–18; Montero 2015. See §4.1 for discussion.
§4.1. The problem of control in skilled action

In this section I briefly (§4.1.1) recount the recent debate, and (§4.1.2) make some clarificatory remarks on the concepts involved.

### 4.1.1. A recent history of the debate over skill

**How reflection controls automatic associations**

The skilled-action debate has been mainly an exchange between those who follow, or seek to improve upon, Dreyfus’ anti-intellectualist theory of skilful coping, and those who object to it. Anti-intellectualists often motivate their case with examples of skill-based actions whose production seems not to require reflection (§2.2.3). Brownstein and Madva, for instance, argue that automatic normativity sufficiently explains, e.g., the museumgoer’s action of standing in the right distance to contemplate a painting. Anti-intellectualists extend this kind of analysis to actions that are too fine-grained and require processes of dynamic adjustment too quick to be performed on the basis of reflective calculation. They think such skill-demanding actions (from playing football, to playing chess, to seeing what is wrong with an architectural design, to finding the proper distance to an interlocutor) are based on intuitive processes of alleviating a felt tension. A similar view is also present in the dual-process psychological tradition. According to a classical account of action and control,

![Deliberate conscious control](image)

> deliberate conscious control is generally agreed to involve serial processing steps, each step taking on the order of 100 msec or more. Such control would simply be too slow to account for skilled human behavior that requires action sequences to be initiated just when environmental or internal conditions call for them; in some situations they must be accurate to the nearest 20 msec. This is consistent with the general view that deliberate control of skilled performance leads to deterioration of performance. (Norman & Shallice 1986, 11)

---

4 Dreyfus (§2.1.2D) has written copiously about the subject. Some representative texts are Dreyfus & Dreyfus (1986); Dreyfus & Kelly (2007); Dreyfus (2002, 2012). Others supporting an anti-intellectualist view of skilled action include Searle (1983, 150–153); Beilock & Carr (2001); Beilock et al. (2004); Ennen (2003); Rietveld (2008, 2010); Klaassen et al. (2010); Beilock (2010); Brownstein & Madva (2012); Di Nucci (2013).


6 Because of the traditional link between reflection and consciousness, Norman and Shallice call this kind of control both deliberate and conscious. I argued above (§3.3) that some kinds of consciousness extend beyond the realm of deliberate attention.
The situated intellectualist can reply that this analysis of automatic behaviour is partially true, but partially false. Partially true, because the museumgoer determines what the right place to contemplate the painting is without performing calculations of distance; but partially—and crucially—false because the associative process that leads her to automatically finding the right place to stand is nevertheless *causally triggered* and *semantically controlled* by a higher-order intention, just like in the case of Yarbus’ eye-saccade experiments ($\S 2.2.3$). The museumgoer’s intention to contemplate the painting governs the automatically controlled sensorimotor processes. Moreover, it is easy to come up with illustrations of automatically controlled behaviour that nevertheless fails to count as intentional action. Think about my driving home out of habit, instead of taking the route I actually intended, and any other case of slips which nevertheless seem to display automatic control (I can skilfully respond to unexpected obstacles on the road and correct for errors), but lack higher-order coordination to intentions and plans. Slips suggest automatic control is not sufficient for (and may even threatening to) intentional action ($\S 3.1.1$).

This intellectualist reply is hard to refute, partly because in most anti-intellectualist examples of motor intentionality (Merleau-Ponty’s organist, Dreyfus’ chess player, Rietveld’s architect, Brownstein and Madva’s museumgoer) agents perform behavioural routines that, though automatically controlled, seem dependent on a concurrent, higher-order intention that structures the process *top-down* by leading attention toward some environmental features that become more salient than the others *because* they are relevant to attaining the higher-order goal. Brownstein and Madva, for one, admit that

> [f]eatures often become salient by virtue of their relation to practical goals and concerns. The features that rise to salience in guiding the museumgoer’s moment-to-moment behavior as she rushes to find a bathroom may differ greatly from those that rise to salience when she is ambling aimlessly. (2012, 420–1)

This co-variance of salience and higher-order intentions should concern the anti-intellectualist. It suggests that in each practical situation environmental features are rendered salient *in response to* the causal and semantic control exercised by higher-order states. And if this is so, then the normativity of automatic-associative processes can easily be construed as subordinated to the normativity of reflection.

---

7 See Wu (2011) for a detailed intellectualist account of how intentions govern automatic processes by structuring attention.
Anti-intellectualists may reply that the reflective processes involved “don’t control the subsidiary coping movements; they merely cause them. That is, just as the coping skills don’t play a role in planning, cognitive planning operations don’t play any role in coping” (Dreyfus 2002b, 422). But, as pointed out above on the basis of Yarbus’ experiments (§2.2.3), the situated intellectualist offers evidence that associations are cognitively penetrable and therefore sensitive to the semantic properties of reflective states. 8

**Three psychological phenomena against intellectualism**

Some anti-intellectualists have recently played a different card: they have appealed to certain psychological phenomena of expert performance that do not seem to be susceptible to intellectualist explanation. 9

**(A) Reflection, choking, and the yips**

Some intellectualists hold that automatic behaviour constitutes intentional action only if the agent can answer ‘Anscombean’ questions (such as ‘What did you do?’ and ‘Why did you do that?’) about it. 10 Against this view, Brownstein (2014) invokes a psychological phenomenon in sports variously called “choking”, “the yips”, and “Steve Blass disease”:

Steve Blass was a hugely successful Major League Baseball player who inexplicably lost his ability to pitch. Within the span of 3 years he went from winning 20 games in a season and pitching Game 7 of the World Series to playing in the minor leagues, and shortly after that to being out of professional baseball altogether. It wasn’t just that Blass lost his edge. Suddenly and inexplicably, he couldn’t even hit the catcher’s mitt from the mound. There was nothing physically wrong with Blass, so far as anyone could tell. The common interpretation is that Blass’ problem stemmed from something like chronic over-thinking. This is a common enough “disease” in profession sports to earn its own name. (Brownstein 2014, 557) 11

---

8 See Fridland’s (2014; 2015) argument that automatic processes are cognitively penetrable because the semantic content of intentions can inform them by structuring attention. “Thinking about attention in this way, we can see why, e.g., a baseball player who has decided to steal second base will automatically shift his focus to certain relevant features of the pitcher’s delivery whereas the player who has decided to wait for the a hit to advance, will perceptually attend, automatically, to different relevant features of the play.” (Fridland 2014, 2747).

9 Sutton (2007); Di Nucci (2013); and especially Brownstein (2014) discuss these anti-intellectualist arguments.

10 See Velleman (2008); Railton (2009); Annas (2011).

11 See also Dreyfus (2012). The phenomenon is known in different contexts under a wide variety of names (also including “freezing,” “the waggles,” “the staggers,” “the jerks,” and “whiskey fingers”). Unsurprisingly, these terms serve as catch-all phrases for what seem to be several different conditions, still not fully understood (for an overview see Owen 2014).
Brownstein argues that reflective self-awareness of the kind that would be necessary to answer Anscombean questions is precisely the kind of over-thinking that leads to the yips. Multiple studies seem to confirm Brownstein’s view that self-centred reflection is harmful to skilful performance.\textsuperscript{12} This in turn lends support to the anti-intellectualist view that reflection plays no role in accounting for the fine-grained sensorimotor control displayed in expert performance.\textsuperscript{13}

\textit{(B) Expertise-induced amnesia}

Second, evidence suggests that the automation necessary to acquire a skill generates a kind of “expertise-induced amnesia”, a disconnection between what is done and what is accessible to self-report.\textsuperscript{14} Many expert testimonies seem to confirm this. Take basketball star Larry Bird: “[a lot of the] things I do on the court are just reactions to situations […] A lot of times, I’ve passed the basketball and not realized I’ve passed it until a moment or so later”; Hall of Fame NFL running back Walter Payton: “people ask me about this move or that move, but I don’t know why I did something. I just did it”; and baseball catcher Yogi Berra: “Think? How can you hit and think at the same time?”\textsuperscript{15}

For the anti-intellectualist, expertise-induced amnesia speaks against intellectualism about expert action. If a behaviour’s production implies a reflective process, some aspects of what she did and how she did it must have gone through working memory; and if so, the agent should be able to recall and report some of those aspects (since reportability and working memory seem to be linked); the fact that experts have a hard time reporting \textit{what} and \textit{how} suggests that their actions is produced by automated, working-memory-independent associative processes. So expertise-induced amnesia supports the view that automatic (rather than reflective) control is what accounts for expert actions.

\textsuperscript{12} E.g. expert golfers who have been describing their technique need twice as many attempts to sink a putt than those who had previously performed an unrelated task (Flegal & Anderson 2008); and expert golfers were more accurate when their attention was diverted from their own performance than when they were instructed to focus on it, while novice golfers presented the opposite accuracy pattern (Beilock et al. 2004).

\textsuperscript{13} Taking phenomena like this as evidence, Di Nucci (2013) has argued for the extreme intellectualist view that a large portion of our actions is entirely “mindless”.

\textsuperscript{14} The term was coined by Beilock and Carr (2001).

\textsuperscript{15} Bird is quoted by Dreyfus & Dreyfus (1991); Payton by Beilock (2010, 224); and Berra by Sutton et al. (2011, 80). Brownstein (2014, 555–6) reports these and several more expert testimonies indicative of expertise-induced amnesia.
(C) Expert confabulation
Evidence for expertise-induced amnesia notwithstanding, some experts do produce reports about what they do and how they do it. The anti-intellectualist would hypothesize that those reports should often be mistaken, consisting in post-hoc rationalizations rather than direct reports of the action-production processes. And indeed the reports seem to be often inaccurate. Several studies suggest that the rules experts seem to reflectively believe to be following (about e.g. how and where to focus attention) turn out to misdescribe what they actually do. Cricket batters, for instance, are taught that they should always keep their eyes on the ball, and experts seem to believe this is what they do. But Land and McLeod (2000) have shown that the more experienced a batter is, the less time her eyes spend focusing on the ball. The expert batter’s eyes move quickly away from the ball (they follow it for only the first hundreds of milliseconds after its release), and make a “predictive saccade” to the place on the ground where they anticipate the ball will bounce. The more expert the player, the faster the saccade; so the expertise is inversely proportional to the time spent watching the ball.

Cases like this seem to support Dreyfus’ account, according to which the performer relies less and less on explicit rules as she becomes more and more experienced. Moreover, the fact that performance does not coincide with the performer’s explicitly held rules suggests that the semantic content of her reflective states and intentions does not (or at least not entirely) control skilful automatic processes. If this is so, then reflective control does not (or at least not entirely) govern automatic control.

4.1.2. On the notion of ‘skill’
So there are three psychological phenomena that seem together to provide quite solid evidence in favour of anti-intellectualism about skilled action. Before moving on to assessing whether this is indeed the case, allow me to make a couple of conceptual clarifications.

Skill and automaticity
What do concepts like ‘skill’, ‘skilful’ and the like mean? The precise nature of skill is what is at stake here, but for now I posit that having a skill implies at least having acquired through practice a certain level of automaticity relative to the performance of an activity’s cognitive or sensorimotor sub-routines. Thus, doing Φ skilfully implies that certain components of Φ-ing have become automated, or ‘chunked’, and are thus performed without involving working memory. (E.g. typing these words skilfully implies having automated the processes that associate
specific letter strings with specific finger movements, so that I no longer need to reflectively think about how to move my fingers in order to write each specific word).

Automaticity is thus necessary but for skill. But it does not seem to be sufficient, because not all associations we have developed would qualify as skills (e.g. those that make up our implicit biases). What else does skill need, then?

**Skill and control**

It seems that in order to count as a skill, and not merely a bunch of automated processes, at least an element of control is required that unites automatic routines into a coherent performance. What distinguishes a novice from a professional, and a professional from a world-class expert, is that in performance the latter displays an extraordinary level of fine-grained control with respect to every one of her skill’s multiple aspects. Experts are able to fine-tune each of their body movements to the particular situation, direct their attention to exactly the relevant features while inhibiting the rest, and make multiple precise corrections on the fly, all of this within a coherent, unified skilled performance.

But what kind of control is this, that separates the amateur from the professional, and the professional from the expert? That is precisely the issue. Anti-intellectualists hold that only automatic control can account for such fine-grained, fast-paced control over perception, attention, and movement. And this seems intuitively correct: after all, reflection is effortful, slow, and capacity-limited, so if we had to perform many reflective computations to supervise the multiple simultaneous automatic processes involved in a single performance (e.g. the right place to stand, where to look, and with how much force and at what angle to strike the ball) we would probably choke. However, intellectualists hold that this cannot be the whole story, because a collection of automatic-associative processes cannot explain the coherence of skilful control. From this perspective, anti-intellectualists like Dreyfus give up on explaining control in skill rather than account for it: they just hold that the true nitty-gritty control happens at the level of multiple subpersonal, over-practiced associative processes, and that at the personal level the agent must simply not think too much, or not think at all—just do, instead of think. Thus

---

16 See Logan (1985). Logan claims that a further difference between skilled and non-skilled agents, besides automatized sub-routines, is that the former have much more craft-related explicit knowledge (e.g. a skilled rock guitarist will know a lot more about guitar maintenance, amplifiers and effect pedals, and rock guitar legends, than a less skilled guitarist). However, he does not go as far as to explain how explicit knowledge and automated sub-routines are supposed to interact in skilled performance. Further, there may be exceptions to the explicit-knowledge criterion; consider punk rockers who could not care less about explicit musical knowledge, but are nevertheless skilled at punk music performance.

17 See Fridland (2014).
anti-intellectualism turns skilled control into a passive, sub-personal phenomenon, in which the person as a whole must simply get out of the way.\textsuperscript{18}

But then, if agents must get cognitively involved in performance at the personal level, would this not introduce some ‘reflective distance’ between performer and performance, ultimately leading to performance failures like the yips?

So it seems clear that the kind of control specific to skilled action (which unifies the multiple threads of automatic processing together into a coherent performance) remains unaccounted for. What is unclear is exactly what this control amounts to.

\textit{Kinds of skills}

The terms ‘skill’ and ‘skilled action’ seem quite ambiguous. It is rather unlikely that there is a common essence shared by the skills involved in my tying my shoelaces, an expert chess player’s selecting a move, a cricket batter’s batting, a physician’s producing a complex diagnosis, and a philosopher’s critically assessing an argument. With this worry in mind, I will focus in this Chapter on what could be loosely called ‘sensorimotor skills’, i.e. those involved in mostly bodily performances, like sports and dance. (In the next Chapter I will have more to say about the cognitive skills involved in decision making.)

\textit{Skill and expertise}

One last preliminary remark: some researchers understand skill as a continuum between novices and experts, and thus conceive of expertise as nothing but the highest degree of a skill.\textsuperscript{19} Others trace a qualitative difference between skill and expertise. One way to draw this qualitative distinction is by saying general skill can be constituted of entirely automatic dispositions (and may thus succumb to stagnation), but expertise is reflectively guided toward continuous improvement.\textsuperscript{20} Situated intellectualists tend to support this kind of qualitative distinction, whereas anti-intellectualists lean towards a quantitative distinction, with expertise being fully automatic skill. I will not presuppose either of them, but rather assess the arguments from both perspectives when necessary.

\textsuperscript{18} For this line of argument see Montero (2010); Fridland (2014).
\textsuperscript{19} E.g. Logan (1985); Brownstein (2014).
\textsuperscript{20} A distinction along these lines is made between Annas (2011) and Montero (2015). Papineau (2014) proposes an analogous difference between “practice” (where the performer may automatically go along with the motions) and “competitive performance” (where the performer must keep her reflective focus on the task to seek any possible advantage).
§4.2. Does reflection lead to choking?

For Brownstein, “Steve Blass Disease is an extreme form of a common problem. The problem is that performing skilled actions well stands in an inverse relationship to self-focused awareness” (2014, 557). This is why Steve Blass Disease (or choking, or the yips) offers evidence against Anscombean intellectualism: the performer’s reflectively focusing her attention on her ongoing behaviour—in a way that would be necessary to answer Anscombean questions—hinders skilful performance.

In this section I will assess whether the phenomena do actually support anti-intellectualism about skilled action. I will first (§4.2.1) clarify the meaning of the terms ‘choking’ and ‘the yips’ along with their likely causes; and then (§4.2.2) assess whether the terms, more clearly defined, do suggest that reflecting on what one is doing is harmful to skilled performance.

4.2.1. Clarifying the terms

The issue needs careful conceptual analysis. For on the one hand, it is unclear what we mean exactly by ‘choking’ and ‘the yips’, whether they are the same thing, different phenomena with common underlying causes, or independent situations. And on the other hand, if the phenomena have to do with reflective over-thinking, it is not clear what it is that people are reflectively thinking about when they suffer from these problems, and whether the effect occurs when people reflect on some features but not when on others. It would be useful to query the literature with these doubts in mind.

Performance failure under pressure: one or many?

First, let us deal with the notion of ‘choking’ itself. In the psychological use of the term, an agent *chokes* whenever she performs below what is expected, given her skill level, in high-pressure situations, i.e., in situations where the agent feels that performing well is highly important (Baumeister 1984, Beilock & Gray 2007). Colombians from my generation are quite familiar with this phenomenon. In the eve of the 1994 World Cup, the Colombian soccer team was among the competition’s favourites: they had a brilliant set of players at the peak of their careers, and showed top-level performances in the qualifying games, famously beating the mighty Argentina 5–0. When the moment of truth came, they failed. Actually, ‘failed’ does not do it justice. They flopped horrendously, losing all games but one and coming back home after a first-round elimination. How to explain what happened to them?
Well, there are two main theories about why skilled performers choke under pressure: the explicit monitoring theory holds that high pressure leads the agent to closely monitor aspects of performance that have been automated through practice; this reflective focus disrupts automatic control, and ultimately leads to a performance breakdown. The distraction theory, on the other hand, suggests that high pressure causes the agent to reflect on aspects extraneous to her task (e.g., the outcomes that are at stake); such extraneous thoughts thus turn a one-task environment into a dual-task situation, occupy attentional and working-memory resources, and thus ultimately disrupt performance. The two theories point to two different roles of reflection in choking, explicit-monitoring theory being congenial to anti-intellectualism (reflection on the task leads to choking), distraction theory to intellectualism (reflection on the task is required to avoid choking). So which one is right?

Both are, apparently. The two different kinds of pressure seem to have two different effects on reflection, which in turn seem to cause two different situations of performance failure. DeCaro et al. (2011) showed this by distinguishing, on the one hand, between the pressure—the pressure of being monitored by others and the pressure of high-stakes performance-dependent outcomes—and, also distinguishing two kinds of tasks—reflective tasks that rely heavily on working memory, and intuitive tasks that do not. The researchers found that one kind of pressure affects the performance of one kind of task and not the performance of the other kind. Specifically, the pressure of being monitored hinders the performance of intuitive tasks, and not the performance of reflective tasks (which supports explicit-monitoring theory), whereas outcome-related pressure has the inverse effect: it disrupts the performance of reflective tasks, but not the performance of intuitive tasks (which supports distraction theory). DeCaro and colleagues thus suggest that the pressure of knowing you are being watched can lead you to try to reflectively control automated aspects of your performance, thus hindering automatic control over your action (as explicit-monitoring theory would suggest), while outcome pressure can lead you to reflectively focus on things other than your task, thus also hindering your performance (as distraction theory would hold). Of course, both kinds of pressure can occur together, like when you are playing at the World Cup—and in those cases both kinds of performance failure can occur simultaneously. As many Colombians of my generation know.

It follows from DeCaro et al.’s research that the phenomena anti-intellectualists refer to in support of their view actually consists in at least two distinct conditions. Both conditions are failures of performance under pressure, but one condition occurs when reflection hinders

---

21 See Beilock & Carr (2001); DeCaro et al. (2011); Montero (2015).
automatic control, while the other condition occurs when reflection hinders reflective control. Papineau (2014) helpfully suggests that, although both phenomena are instances of a general phenomenon (which he calls “Not Having Your Mind Right”), they should receive different names and distinct explanations. I will follow him in using “The Yips” to refer to cases in which performance failure occurs due to reflecting on automated components of an action, and using “Choking” to refer to cases in which performance failure occurs due to the agent diverting reflective focus away from performance. (I also follow him in capitalizing the names when I intend them in the narrower, more precise meaning just defined.)

Crucially, Choking and The Yips differ in their outcomes (in one of them reflective control is compromised; in the other automatic control is), and they also differ in their aetiology. Performers who suffer from The Yips lose control because they direct their attention to automated sub-routines constitutive of their skilled actions, whereas those who suffer from Choking lose control because they direct their attention away from their skilled actions.

**Choking, The Yips, and reflective control’s proper role**

From this description of Choking and The Yips it seems to follow that reflecting on the action components is harmful, but not reflecting on the action itself also is. So, what is the proper role of reflection in skilled action? Papineau (2014) draws another distinction, that turns out to be useful here, between basic actions (i.e. behaviours that you can directly decide to perform without having to decide to do anything else), and the components of basic actions (i.e. simpler behavioural routines that are constitutive parts of a basic action and have become automated through practice). I can, for instance, perform the basic action of riding my bicycle down to the bakery, whereas the sub-routines that make up that action (balancing, pedalling, steering, making eye saccades, and so on) are automatically controlled and do not themselves constitute basic actions, since I did not make a decision to do them. Of course, what counts as a basic action and what as an action component is relative to the agent’s skills and habits. The crucial point is that, once I have chunked all the bits of behaviour necessary to perform a given basic action (e.g. I have automated each of the sub-components constitutive of bicycle riding), intentionally riding the bicycle would be a very different cognitive process from intentionally performing each of the action’s components. The former leaves the component performance to automatic control, whereas the latter implies reflectively controlling them—something that would be cognitively much more taxing, and would allow for less fine-grained control. Arguably, a skilled bicycle rider would be much worse at reflectively balancing, pedalling, steering, etc., than at reflectively riding her bike while leaving the sub-routines to automatic control.
So, again, what is the proper role of reflection in skilled action, if reflecting on the action components is harmful, but not reflecting on the action itself also is? The following view seems consistent with the evidence: skilled agents must keep reflection away from action components, but focused on their intended basic actions. For reflecting on action components leads to loss of fine-grained automatic control, and ultimately to Yipping; and losing reflective focus on the basic action leads to distraction, and ultimately to Choking.

The existence of The Yips shows that it is harmful to reflect on automated chunks or action components. Bringing these to reflective focus seems to imply that fast and fine-grained automatic control is overridden by slow, coarse-grained reflective control. This would explain cases of yipping like that of Chuck Knoblauch, the New York Yankees second baseman who became unable to throw to first base, or Mackey Sasser, a catcher who lost his ability to throw the ball back to the pitcher. These are usually taken to be paradigmatic cases of Steve Blass disease (Blass himself became unable to hit the catcher’s mitt from the mound). What they all have in common is that players lose control over tasks that are (1) mundane and simple, and (2) initiated by them, not responses to others’ motions. The Yips therefore affects sportspeople insofar as their function is to initiate plays (e.g. pitchers and catchers in baseball, bowlers in cricket, golfers while putting, or soccer players while shooting penalties), but the condition is notoriously absent from responsive aspects of performance (like batting in baseball and receiving in tennis) and continuous action sequences (like most of soccer).

So The Yips specifically occurs when the performer attempts to control the chunked aspects of her action (things like the specific angle and strength at which to release the ball) reflectively, rather than intuitively (cf. Papineau 2014). One can see how this may be caused by the pressure of being watched: knowing others are watching me can lead me to worry about getting the specific performance parameters right, which would lead me to reflectively check them. It may be advised in these situations to stop worrying about such minutiae and trust one’s instinct.

But Choking is a very different phenomenon. Performers suffer from it when they cannot focus on the task at hand because their attention and working memory are too busy with something else. So the cause of Choking it is a failure to inhibit task-irrelevant features of the situation (like whether I will be able to do it correctly, or what great chances I will lose if I fail). Due to such loss of inhibitory control, the performer’s attention gets crowded by situational features extraneous to the task at hand, and the relevant features get lost in the ‘noise’. What may be advised in these cases is to focus on the performance, to keep your eye on the ball and your head in the game, to stop worrying about the future and be fully present in the situation—not just as a
skilled body that can jump to action automatically, but also as a focused mind that can make
decisions, execute strategies, and shift between tasks in a rapidly changing environment.

4.2.2. Do these phenomena support anti-intellectualism about skill?

It is clear from the above that the only cases that support the anti-intellectualist view of skill are
the cases of The Yips, because reflection on action components can harm skilled performance.
But the cause of Choking is directing reflection away from, rather than towards, the skilled
performance. So do these performance-failure phenomena ultimately support anti-
intellectualism about skilled action? I do not think it does. The Yips does moderate
intellectualism, since it suggests that performers should avoid reflectively controlling action
components while performing (although they may do so while practicing). But Choking gives
support to the situated-intellectualist account of skill, since automatic control of action
components, although necessary, must be subordinated to higher-order reflective processes that
must remain active throughout performance.

This is because Choking reveals that performance can suffer if the agent does not devote
her executive-control resources to keeping her attention focused on the basic action itself. Unless
we do this, we run the risk of not noticing relevant features of the environment that would demand
an adaptation on our part. Everyday driving requires that we keep our mind focused on the road,
to avoid e.g. running the car into a hole in the road, or to avoid other possible obstacles. If this is
so for everyday skilled behaviour, would it not be more so for expert performance?

In this respect, Choking is analogous to mundane cases of slips like when I leave the
office intending to cycle to the candy store, but end up cycling home instead. The behavioural
sub-routines display automatic control, no doubt, but the basic action’s overall unity disappeared
because I did not keep my intention in reflective focus, and this led me to lose reflective control
over it. Analogously, when performers lose focus on their intentions and strategies—perhaps
because menacing and distracting thoughts overload inhibitory control—, this leads to mistakes
during performance, like losing synchronization with teammates or missing some relevant cues
that would have meant a greater advantage had they been noticed. Such mistakes can cost
performers crucial points, or the whole game. But more importantly, sufficient lack of
concentration would imply that the unit that characterizes expert skilled action dissolves into an
pattern of incoherent automatic processes.

The anti-intellectualist argument conflates all cases of performance failure into a single
phenomenon, and explains all of them as caused by excessive performance-directed reflection.
However, on closer inspection it seems that, while reflection on action components can harm performance, reflection on the intended action itself is actually necessary for successful skilled performance. And the more pressure there is on the agent to perform, the more reflective control will be needed. In fact, reflection seems to be necessary particularly to inhibit distracting influences and keep attention focused on relevant features of the situation. The fact that expert-level performance situations are often crowded with distracting forces (like fans of the other team yelling insults at you, fans of your team expecting so much from you, your own worries about whether you will be able to make it or how big of a difference the results will make in your future career, multiple pains around your body, and the general exhaustion caused by performing at your peak for days or weeks) is the crucial reason why simply letting automaticity take control is not sufficient for successfully performing skilled intentional actions at the highest levels.

In conclusion, the phenomena of Choking and The Yips fail to provide decisive evidence for the anti-intellectualist claims that reflective control is unnecessary or harmful for skilled action.

Now, since the situated-intellectualist arguments presented in this section defend the need for reflective control on the basis of the strength of distracting forces, the anti-intellectualist might ask whether reflective control is necessary in situations in which an agent performs a skilled action and no strong distractions are around. In fact, the more expertise is required for successful performance, and the higher the pressure levels are, the need for reflective control becomes more apparent: experts must reflectively direct attention and avoid the multiple powerful distractions in order to perform at their best. But what about everyday, familiar, habitual skilled actions? I would probably need to reflectively control attention in order to win the World Championship of Knot Tying, but do I need reflective control of attention in order to tie my shoelaces? The more habitual and routine the skilled action is, the less clear it is that reflection must be involved.\footnote{In fact, at least some situated intellectualists claim agnosticism about whether reflection’s involvement in habitual actions would be generally detrimental (e.g. Montero 2015). Moreover, Di Nucci (2013, 19) has made the anti-intellectualist argument that the yips affects not only experts, but also agents performing habitual actions: when walking with a glass full of water in your hand, try to reflectively think about how you balance it, and some of that water will be more likely to end up on the floor; reflect on how it is that you write your signature while doing it, and it will probably end up looking sketchy.}

Thus the issue of whether reflection is necessary for skilled \textit{habitual} action remains open, and will be addressed below (§4.5.2). I will first continue to review the other two empirical phenomena cited in support of anti-intellectualism about skilled \textit{expert} action. Up next: experts often claim they ignore how they do what they do. Does this declaration of ignorance point to a lack of reflective accessibility into the workings of expert action?
§4.3. Expertise-induced amnesia

4.3.1. The anti-intellectualist case

Anti-intellectualists hold that the automation of routines implied in skill acquisition causes skilled agents to suffer from “expertise-induced amnesia” with respect to how they perform their actions (Beilock and Carr 2001). Experts themselves often report such amnesia (§4.1.1B), and inquisitive fans could also attest to its existence when, after a great game, their sports heroes are interviewed and have nothing minimally insightful to say about how they achieved their prowess (Sutton 2007, 769).

The theoretical model behind expertise-induced amnesia is rather simple (Beilock & Carr 2001, 703). It stems from the view that working memory can take as input only the information that is broadcast into global workspace (on which see §3.2.2). This implies that anything accessible by working memory must also be available for subsequent declarative report. So, given that experts often fail to report what they do or how they do it, the cognitive processes that determine their performance must have occurred outside of working memory. And given the link between working memory and reflection, the lack of reportability suggests skilled action is controlled automatically rather than reflectively.

The anti-intellectualist can also point to suggestive empirical evidence. Beilock et al. (2002, exp. 1) studied the performance of novice and expert golfers while putting with a normal putter vs. while putting with an S-shaped “funny putter” (see Fig. 4.1). The funny putter was designed to function as a novel constraint that would invalidate experts’ overpracticed sensorimotor routines.23 To assess performance reportability, each participant was asked to take a break from putting and list all the steps composing a typical putt; and once they were done putting they were asked to list all the specific steps they themselves took in performing their last putt.24

Experts gave overall longer generic descriptions of the steps involved in putting than novices (which corroborates Logan’s (1985) idea that skill implies explicit knowledge). But the experts’ generic descriptions were much longer and more precise than their episodic descriptions of

23 Beilock and her colleagues define the contrast between novices and experts differently in several experiments. In this case, novices were participants with no previous golf experience, and experts were “local high school and college students with 2 or more years of high-school varsity golf experience or a Professional Golfers’ Association (PGA) handicap less than 8”.

24 To assess the former, they asked: “Certain steps are involved in executing a golf putt. Please list as many steps that you can think of, in the right order, which are involved in a typical golf putt”. To assess the latter, they asked: “Pretend that your friend just walked into the room. Describe the last putt you took, in enough detail so that your friend could duplicate that last putt you just took in detail, doing it just like you did.”
their own putts. In fact, the episodic descriptions of regular-putter experts were the shortest of all episodic descriptions. This provides evidence for the existence of expertise-induced amnesia, by showing that, although experts have a large store of general explicit knowledge about their skill, what exactly they do while performing is less accessible for them than for novices.

Moreover, experts putting with the funny putters provided the longest episodic recollections of all groups. This suggests that, when experts perform in an unfamiliar situation that forces them to reflectively focus on their task, they can access what they do, and how they do it, in more detail than novices and experts who act naturally. —Of course, the quality of funny-puter experts’ performance was worse than that of normal-puter experts.

Figure 2.1: Beilock et al.’s (2002) funny putter

So the evidence for expertise-induced amnesia seems to support the anti-intellectualist claim that online skilled-action control works independently from working memory.

4.3.2. Intellectualist replies
Intellectualists remain unconvinced by these two arguments—i.e. the argument from experts’ anecdotal reports, and the on from funny-puter-type studies—, and have provided replies to both. Let us start with a reply to the latter.

The experiments are largely inconclusive
It is worth noting, first, that Beilock’s studies do seem to speak against Anscombean intellectualism. In fact, they seem to show that more highly skilled performers are less able to reflectively access
information about the *what* and *how* of skilled performance (this as measured by their ability to provide a list of the steps taken to perform the skilled action) than novices. Having said that, this is not sufficient proof to reject other kinds of intellectualism, which may claim reflection’s role in skilled action production is different from specifying the *what* and *how* of performance.

What other role might reflection play in skilled action production? One candidate for reflection’s role (already mentioned in §4.2.2) is sustaining attention on the current task, so that relevant environmental features may become salient, and inhibiting powerful irrelevant stimuli from capturing attention. In a word, reflection may play the role of executive attention control. It seems that expertise-induced amnesia of the kind presented in Beilock et al. (2002) does not provide sufficient evidence to put this kind of intellectualism into question.\(^{25}\)

On the other hand, Montero (2015) has argued that there is a further problem with taking Beilock’s experimental findings as evidence for the nature of expert action, a problem lying in the very definition of ‘expertise’. Beilock’s studies use a gradual conception of the skill-expertise distinction, so they use ‘experts’ to refer to the group of the most skilled or experienced participants. But this begs the important question whether experts are just *highly skilled agents*, or whether expertise is a qualitatively different state from skill.

Montero defends the latter. In her view, experts are “those who have practiced their skill in a thoughtful, critical manner on a near daily basis for at least around 10 years and are still practicing their s[k]ills in such a way” (2014, 386), so the group of experts largely coincides with those we consider *professionals*. Understood this way, Beilock’s experiments do not really show anything about the nature of expertise, because none of its participants are real experts. Maybe if professional experts participated in the funny-putter experiment, they would be able to produce detailed descriptions the steps they took to produce their putt.\(^{26}\) Maybe they would not; but the general point is that Beilock’s experiments do not provide evidence one way or the other.

In sum, the anti-intellectualist case based on empirical evidence for expertise-induced amnesia seems inconclusive.

*Alternative expert reports suggest otherwise*

What about the argument from expert reports? Being merely anecdotal, it cannot be very solid on its own. But it looks even less solid when contrasted with *other* reports in which *professional* experts

\(^{25}\) To be fair, Anscombean intellectualism was the target of Brownstein’s (2014) argument, so it is not a fault of the argument that it does not do more than it was designed to do.

\(^{26}\) This would be in accordance with Montero’s view that experts are characterized “by engaging their conscious minds during analytical, thoughtful and effortful practice” (2015, 387).
(of the kind even Montero would acknowledge as such) seem what their minds do while they perform at the top of their game. Studies by Nicholls et al. (2006) and Oudejans et al. (2011) used retrospective methods (like diaries and verbal recollection) to find out what thoughts professional athletes from a wide variety of sports had while performing under pressure. When asked what they did to cope with pressure, it was found that “the most effective coping strategies that were used on a frequent basis were increasing concentration on task and increasing effort” (Nicholls et al. 2006). It was also found (Oudejans et al. 2011) that in high-pressure situations athletes report relying on “positive monitoring” strategies (some of which they describe as “I focus on what I should do” and “I make sure that I work harder”) to maintain high-quality performance.

To the extent that expert athletes themselves report these two methods as highly effective in coping with pressure, they seem to counter the expert reports, cited by anti-intellectualists, that they do not know why or how they do what they do. It is likely that experts are unaware of why and how they structure their performances the way that they do (probably this structuring process is largely due to automatic responses to the practical environment’s solicitations), but one thing that experts seem to be aware of is that in order to properly respond to the environment’s solicitations — especially in high-pressure situations — they need to stay focused on the game (which includes inhibiting negative, task-external thoughts) and engage effortfully in the action.

This argument is, of course, based on anecdotal and introspective evidence, so it is as limited as this type of evidence is. But its main purpose here is to counter the intuitive force of the statements of Larry Bird and the others, as cited by anti-intellectualists; and it seems to succeed at it. For, to the extent that these self-reports are right, they go against the anti-intellectualist view of skilled action as a cognitively effortless responding to the situation’s solicitations. Expert agents are aware they cannot be mentally passive and just let the subpersonal processes fire away in a state of effortless flow. On the contrary, experts describe skilled performance as effortful both physically and mentally. This suggests that expert-skilled action requires reflective control, insofar as it requires a considerable measure of executive control of attention. Experts may not be aware of why and how they structure their movements during performance; but they do seem aware that in order to be able to do it, they must keep their focus on the task.

So the arguments from expertise-induced amnesia seem to have two negative upshots. One of them is that both expert reports an empirical evidence make Anscombean intellectualism implausible. The other is that, despite this, expertise-induced amnesia does not provide conclusive evidence for anti-intellectualism about skilled action, and rather offers partial corroboration for the attention-structuring intellectualism that the analysis of Choking and The Yips had already suggested.
§4.4. Expert confabulation

4.4.1. Evidence for expert confabulation

Not all experts remain silent when asked what they do and how they do it. When they describe what they do in performance, they often mention rules that they claim to follow. But some empirical evidence suggests the rules they mention tend to misdescribe what they actually do.

Some of the strongest evidence for expert confabulation comes from Land & McLeod’s (2000) study of cricket batters, which shows that the time batters spend watching the ball diminishes as skill level increases (§4.1.1C). This entails that the popular ‘Keep your eyes on the ball’ rule—which many batters seem to endorse—does not match semantically with what skilled performers do. Instead of following the ball, skilled performers seem to perform predictive saccades to move the gaze ahead of the target, or catch up with it.27 Moreover, cricket training makes special emphasis on the importance of watching the ball right at the precise time of hitting it.28 But this seems particularly implausible, given the high speed at which the ball travels. Particularly, Land & McLeod (2000, 1342) claim that batters track the ball during the first 0.2 seconds after it is released, leave it behind to make the predictive saccade, then loosely follow it as it approaches the bat, but ultimately lose it from sight near the end of its trajectory.29 Moreover, the predictive saccade makes functional sense, in that it seems to help the batter predict the ball’s future trajectory, thus making a contact with it more likely. So experts seem to be wrong both in that they keep their eyes on the ball, in that they do so at the crucial time of striking it, and even in that it is normatively relevant to do it. In short, it seems that “[c]onfabulation is common when skilled agents are asked to report the techniques they use ‘on the field’ ” (Brownstein 2014, 560).

---

27 Moreover, there is evidence that this (i.e. using predictive saccades instead of directly following the ball) occurs not only in cricket (Land & McLeod 2000; Croft et al. 2010; Mann et al. 2013), but also in baseball (Hubbard & Seng 1954; Bahill & LaRitz 1984), table tennis (Ripoll & Fleurance 1988), squash (Hayhoe et al. 2012), and soccer (Savelsbergh et al. 2002; Savelsbergh et al. 2005) performers. There is also evidence that catchers misreport the cues they use to determine where to stand in order to catch the ball (Reed et al. 2010). For discussion see Koedijker & Mann (2015).

28 Mann et al. (2013) report Justin Langer (retired batter and coach for the Australian team) claiming that the idea of not watching the ball while hitting it was “unbelievable”, and one of the world’s all-time scorers holding that he always aims at seeing the ball “come from underneath my bat” as he hits it. At the same time, Mann et al. (2013) also report several studies suggesting that achieving these feats would be rather impossible, given the speed at which the ball travels.

29 See also Bahill & LaRitz (1984); Ripoll & Fleurance (1988).
4.4.2. “Expert”? “Confabulation”?

That said, the study by Land and McLeod seems susceptible to the same kind of terminological objection as those of Beilock and colleagues: ‘experts’ are roughly understood as ‘those participants with the highest degree of skill’.30 This seems to assume that there is a merely quantitative difference between skilled and expert performers. But would this be valid also for true experts, i.e. top performers?

Mann et al. (2013) tested this by comparing elite performers (among those with the highest historical records) and highly-skilled club-level performers (who play recreationally at local cricket clubs). They found a couple of relevant features proper to elite batters’ performances: first, elite batters consistently follow the ball throughout the whole trajectory with their heads, even if not with their eyes. Second, elite batters make not one, but two predictive saccades: the first one toward the place where the ball will bounce, and the other toward the place where the ball will make contact with the bat. This second saccade gives elite performers the capacity to actually see the ball as they hit it, an ability that is much more reliable in them than in club-level performers. Mann et al. compare elite batters’ capacity to follow the ball with their heads to a “miner’s torch” (2013, 6): if elite batters had one, the light from the their torch would always illuminate the ball; not so for the club-level performers. They think that this capacity to track the ball’s movement with the head gives elite players an advantage, because it simplifies the cognitive problem of establishing the ball’s trajectory. The ball remains stable in egocentric space (i.e. it is in a constant position relative to the head); so, since they already know where (in egocentric space) the ball will be, all they need to figure out is when the ball will reach them.

The ‘when’ seems to be specified by the second predictive saccade. In fact, a crucial difference between the two groups is that elite batters moved their eyes so that they were always at the ball or ahead of the ball, whereas the club-level batters’ eyes sometimes lagged behind the ball, and had to make saccades to catch up with it. This suggests that the saccades deserve the name ‘predictive’ more for the elite performers but not quite for the club-level performers. So the elite batters’ saccades have a truly predictive function (as opposed to the catching-up saccades of club-level performers), in the sense that they anticipate the key places where the ball will be.

What does this mean for the discussion about expert confabulation?

Land and McLeod drew several conclusions from their study: skilled batters in general did not keep their eyes on the ball; they often failed to see the ball as it was hit; and there was no

---

30 Land & McLeod’s participants were “Mark, a professional cricketer who has opened the batting for Warwickshire, Charlie, a successful amateur who plays Minor Counties cricket for Oxfordshire, and Richard, an enthusiastic but incompetent amateur who plays low-level club cricket” (2000, 1345).
“systematic difference” in eye movement between skill levels (2000, 3142). All of this was used by the anti-intellectualist as support for the phenomenon of expert confabulation. But Mann et al.’s study is good news for intellectualists, for several reasons. First, it makes explicit coaching advice and rules much more consistent with actual practice than anti-intellectualists (and Land & McLeod’s study) had suggested. In fact, truly expert batters do watch the ball as they hit it, and they do keep their eyes on the ball throughout its trajectory, in the sense that the ball is always in the visual field determined by their head’s position (so it is at least peripherally visible at any time). This ultimately raises general doubts about the evidence quoted as support for expert confabulation, by suggesting that the people categorized as experts in those studies may just be skilled performers, rather than true experts.

Moreover, the study suggests that there is (at least in this case) a true qualitative difference between real experts and merely skilled agents, as suggested by the facts that elite performers make two saccades, whereas skilled players make only one, and only elite performers can watch the ball as they hit it. One can then say, with the intellectualist, that perhaps the reason skilled players fail to watch the ball is that they are not skilled enough. If this is the case, then rules like ‘Keep your eyes on the ball’ can be interpreted as normative rather than descriptive, and failures to comply as failures of performance.

So there are reasons to doubt that the evidence for expert confabulation actually is evidence of what experts do, or evidence of cases of confabulation. The phenomenon therefore seems to be amenable to intellectualist explanation.
§4.5. Reflection’s role in skilled action

This examination of the psychological evidence about Choking and The Yips, expertise-induced amnesia, and expert confabulation does not seem to have left many results in favour of anti-intellectualism. It has, on the other hand, provided some reasons to suspect that reflection is more fundamentally involved in skilled action than may have been originally thought. In this section I will (§4.5.1) seek to clarify the role of reflection in expert-skilled action, in accordance with the conclusions from the previous sections; and (4.5.2) tackle the question whether reflection plays any role in the performance of habitual and mundane skilled actions.

4.5.1. Reflection as the intention-based modulation of attention

First, a quick summary of relevant data. I have argued that (§4.2) the phenomenon of The Yips does show that reflecting on the automated sub-components of action is harmful to performance, but the phenomenon of Choking suggests that allowing reflection to drift away from the basic action itself also harms performance. So personal-level reflection seems to play a sort of unifying role for the different sub-routines that constitute a skilled performance. I have also argued that (§4.3), although experts claim they have no clear idea about what they do or how they do it, they do seem to be clear about the fact that in order to do it they must make efforts to keep focus on their task, especially in moments of pressure. So although expertise generates amnesia about the what and the how of performance (which speaks against an Anscombean kind of intellectualism), it does not render agents oblivious of the fact that task-oriented focus is necessary for performance. This last point suggests that reflection may play a kind of attention-structuring role in expert action. Finally, I have argued that (§4.4) the evidence from real, world-class experts does not support expert confabulation.

The studies of cricket batters do suggest, however, that the way in which visual attention is structured varies in accordance with levels of skill and expertise: Eye-saccade studies (like Land & McLeod 2000) show that more highly-skilled performers make faster predictive saccades and focus on more relevant aspects of the environment; and Mann et al. (2013) showed that elite experts make a crucial second predictive saccade that enables them to reliably look at the ball as they hit it. If this is so, then expertise and skill seem to be intrinsically linked with the ability to structure attention, and different levels of skill and expertise are related to different ways to structure attention. This supports the view mentioned earlier (§2.2.3) that reflection’s
task in skilled-action production is the intention-based modulation of attention.\footnote{This is a view recently developed particularly by Wu (2011, 2013a, 2015), and also by Fridland (2014, 2015) and Shepherd (2014, 2015). In this subsection I follow Wu's line of argument; and in the next one I propose that it should be extended to include even the most habitual intentional actions.} It is time to present this view in some more detail.

Consider this example: When I play soccer with my good friend Lucho, a great number of the situation’s aspects could activate different automatic-associative behaviours. The situation is crowded with affordances (i.e. with possibilities and invitations for action): seeing my friend in the field affords reaching and chatting with him. But he happens to be on the opposite team, and just scored a goal, so noticing him around also elicits anger and affords punching him. The soreness in my left foot affords taking the shoe off and staying still for a while. And while I’m at it, the nice grass underfoot may also afford tumbling on it and enjoying the sun. My soccer skills (assuming I have any) also afford a bunch of potential moves conducive to playing a good match.

Of all these possible sensory, affective, and motor affordances, how do I distinguish what is relevant from what is irrelevant in the moment? The crucial lesson from sections §4.2–4.4 is that achieving an appropriate distinction between relevant and irrelevant affordances—one that allows me to grasp the tiny details that distinguish a decent performance from a great one—is very unlikely without the top-down modulation of attention on the basis of an occurrent goal. I can only focus on the soccer-based affordances if I manage to inhibit the chatting, punching, resting, etc. affordances; but doing this implies inhibitory control, which as mentioned before is one of the executive functions of working memory (broadly construed) (§3.2.2). Only through this executive, top-down control of attention can the relevant features of the situation acquire the special glow, the mandatory character that “carries me forward as if by magic” (to use Csikszentmihalyi’s expression), or that “summons” the correct response from me (to use Dreyfus’ term). If attention is not hierarchically structured by this goal, all affordances would remain equally attractive, which means that no affordance would really be all that attractive. Moreover, achieving this distinction between what is and what is not relevant, so that the relevant carries me forward, becomes more demanding the higher the pressure level, because when the pressure is higher the task-irrelevant features have a greater intuitive salience, and therefore inhibitory control (a limited capacity, like all executive functions) has a greater role to play.

From this perspective, it seems that anti-intellectualist descriptions of skilled action are incomplete. Take Merleau-Ponty’s classical description of the way a player experiences the field:
For the player in action the football field [...] is pervaded with lines of force (the ‘yard lines’; those which demarcate the ‘penalty area’) and articulated in sectors (for example, the ‘openings’ between the adversaries) which call for a certain mode of action and which initiate and guide the action as if the player were unaware of it. [...] The player becomes one with [the field] and feels the direction of the ‘goal’, for example, just as immediately as the vertical and the horizontal planes of his own body. [...] Each maneuver undertaken by the player modifies the character of the field and establishes in it new lines of force in which the action in turn unfolds and is accomplished, again altering the phenomenal field. (1942/1965, 168–169)

The problem with this description is that it includes only the relevant lines of force: the ones that are task-relevant. But the field is usually pervaded with way too many lines of force, which call for way too many responses that could guide behaviour in all sorts of incoherent directions. Automatic control over the action’s subroutines is necessary to explain why the fields of force guide the action as if the player were unaware. It also explains why experts cannot report the what and the how of performance (since it is largely structured sub-personally as automatic responses to environmental features). But automatic control does not seem able to explain why only the relevant lines of force guide skilled action, while the irrelevant lines of force are unattended to or inhibited. The need to perform this inhibitory task is, I think, the reason behind why experts recognize they have to focus on what they are doing. They might otherwise explode in anger, decide to take an untimely break, go yell at some fans of the opposite team, and so on.

So reflection is necessary for expert, skilled action because this mode of action requires the modulation of attention to orchestrate automatic processes (both sensory, cognitive, and motor) toward an intention. In this context, Wu suggests that we conceive of intentions not as mental events that play their causal role before action is produced (like the standard causal theory [§2.2.3] would have it), but rather as structural causes synchronic to action. An intention’s role would be to structure the web of mental associations in such a way that some sensory, affective, and motor associations are primed to be more easily activated, while the activation value of other associations is diminished so that these situational features do not become as salient.

That is, thus, the causal role of intention in skilled action: by modulating the activation levels of automatic associations, an intention structures attention so that the relevant sensory features are primed to become salient, and hooked up with the relevant affective and motor routines. This is how I avoid hugging or punching Lucho while in the field (after the match I
may do both), and instead anticipate his motions, steal the ball from him, and make a crucial pass to one of my teammates.\footnote{By the way, this understanding of intention as modulator of attention is a good candidate for solving the causal deviance problems that plague classical causal theories of action (§2.2.3): from this perspective, causal chains become deviant when the agent fails to reflectively structure attention; this leads to failures in linking sensory and motor behaviours, thus making the subsequent behaviour not count as intentional action (Wu 2015, 11).}

Earlier in the Chapter I mentioned that, although automaticity was a necessary condition for skill, it did not seem to be sufficient, and that what seemed to be lacking was a proper explanation of control (§4.1.2). Along the lines of recent intellectualism,\footnote{Cf. particularly Fridland (2014); Montero (2015); Wu (2015).} I take it that the kind of control that was required to explain skilled action is this kind of top-down, selective modulation of attention that allows for the relevant automatic-associative routines to be activated, and inhibits the irrelevant ones. Now, if that is the role of reflection in skilled action, what is the role of \textit{skill} in this account of skilled action? In other words, what is the difference between the actions of a novice and those of a skilled performer? Well, it turns out that there are two aspects of skill: the automatic routines acquired through practice, and the reflective ability to focus attention on the task-relevant details despite distractions (an ability which I assume performers also perfect through practice). The automatic aspect of skill allows for more relevant behavioural routines to be available to skilled agents: in cycling, once I automate how to keep balance, pedal, and steer at the same time, I can move on to fancier things, like riding with no hands on the handlebar. The attention aspect of skill\footnote{This aspect is arguably not just a matter of reflection. Like cricket batters, it implies sensorimotor skills (like eye saccades) that may be to a great extent automatically controlled.} allows for what the intellectualist has argued since the beginning is crucial to skill itself, and separates the best from the rest: the continuous drive toward improvement, the constant learning process. By reflectively focusing attention on the task-relevant features, the agent can increasingly ‘zero in’ on what is most relevant, and more easily inhibit or disregard the other environmental features. Teaching philosophy gives an opportunity to see this happening live: when teachers face a philosophical text they immediately structure it, attributing various degrees of argumentative and conceptual relevance to each expression and each claim, almost directly seeing the argumentative structure—a part of the philosophical skill that takes a huge amount of practice for new students to learn. And it seems we can continuously improve as we face new texts and discussions.

Now, it may be that truly elite experts are not just quantitatively, but also qualitatively different from the rest of us, in the way virtuous people are different from merely continent people in Aristotle’s ethical theory. (The best among philosophers may be able not only to read better than the rest of us, but may actually be able to do different things with the texts and the
arguments.) This is a matter of empirical investigation, and although there may be some hints around (like Mann et al. (2013) suggest), as far as I can tell we do not yet have sufficient evidence to make a well-grounded judgment.

4.5.2. Fringe actions
If the above is correct, the role of intentions as top-down association modulators is of crucial importance for skilled agents performing at high levels. But what about habitual, everyday actions? The way we tie our shoelaces, ride our bicycles, and make small talk with the neighbours in the elevator, seem to display many skills, even if modest ones. —Do we need more than automatic control to produce such actions?

Are tying shoelaces and catching flying balls intentional actions?
Some intellectualists escape this problem by claiming behaviours like shoelace-tying are too ‘behaviour-y’ to count as intentional actions. But others have a harder time employing that strategy. Bratman, for one, suggests that some intention-independent displays of skill should be considered intentional:

Are there cases of spontaneous activity that, while plausibly classified as intentional, do not involve anything reasonably identifiable as an intention to act? If you unexpectedly throw a ball to me I might reach up and catch it. I catch it intentionally, but perhaps my catching it involves no intention to do something. (Bratman 1984, 395 n. 26)  

Here Bratman approximates the conception of operative and motor intentionality, the intention-independent directedness to the world (§2.1.2A–B). Wu wonders about a similar case:

An expert fencer is standing with her niece in a toy store. At some point, the niece picks up a toy sword and pretends to lunge at her aunt who now automatically, reflexively, perhaps playfully moves in a way that expresses her training in respect of lunges with sharp implements. I don’t think we need to postulate an intention to play with her niece here. The fencer acts skillfully but without explicit intention. (Wu 2015, 20)

35 E.g. Annas (2011) denies that tying one’s shoelaces counts as an intentional action.
36 Montero (2015) also seems to be a part of this group. She rejects “The Maxim”, i.e. the view that reflection interferes with expert skilled action; but she is doubtful about whether to reject or accept “The Restricted Maxim” (i.e. the view that reflection would interfere with “highly automatized, everyday skills” (2015, 112)).
These are examples of skilful, intention-independent sensorimotor behaviour. No reflective control seems to be involved, but the behaviours are clearly intelligent insofar as they are clear displays of automatic control. Moreover, they also seem to be undeniable cases of intentional behaviour, since the automatic control displayed reveals the world-directedness of motor intentionality. The crucial question is: are they also cases of intentional action? If we can find cases of intelligent, automatically-controlled, intention-independent behaviour that nevertheless clearly falls short of intentional action, then we should answer this question in the negative. I argue below that such cases exist.

**Action, automatic control, and the perils of distraction**

The cases I have in mind include the examples of slips mentioned above. Take the story of a person who goes to the porch to take out his car, but once there, as he sees his gardening boots, he puts them on and ends up in the backyard, ready for a gardening session.³⁷ Or the already-mentioned example of making the plan of cycling somewhere out of the habitual, but ending up cycling home instead. Slips like these clearly display control, but we should hesitate to call them 'actions'. To see why, imagine people whose entire behaviour is constituted by slips.

Such people do seem to exist. I am thinking about with a neurological condition called *utilization behaviour*, which is roughly characterized as the appropriate usage of an object in an inappropriate situation.³⁸ Utilization behaviour patients present the following symptoms: they seem to have lost sensitivity to broad aspects of their practical situation, and respond immediately to any of the situation’s affordances. They also seem affectively insensitive to the inappropriateness of their actions. Thus, a patient with utilization behaviour may continuously switch a light on and off; if she sees a comb she may start combing her hair with it, regardless of who owns it; if she is in a bedroom she may undress and go to bed, even if she is just visiting someone else; she may drink from a cup of coffee that is on the desk, although it belongs to her doctor. And after doing any of these things, the patient seems unaffected by the awkwardness generated by her behaviour. A patient may even give Anscombean accounts for her behaviour: she may acknowledge a certain cup of coffee belongs to the physician, and explain that she drank from it because she was thirsty.

---

³⁷ One of Reason & Mycielska’s (1982) participants mentions this happening to him. In their classical discussion of attention and control, Norman & Shallice (1986, 12) take this case as an instance of the behaviours known as “capture errors”. They think these situations provide evidence that “conscious attentional control is not necessary for the initiation or execution of action sequences”. It is unclear, however, how rigorously they mean for this behaviour to count as ‘action’.

³⁸ Pandey & Sarma (2015). The condition was initially reported by L’Hermitte (1983). The condition is linked with bilateral damage to cortical and subcortical locations in the medial premotor system. For discussion of the syndrome and its relevance to agency and control, see Baddeley (2007, Ch. 17); Rietveld (2012).
In his discussion of utilization behaviour, Rietveld (2012) holds that patients, while sensitive to affordances, are insensitive to the comparative relevance of different affordances. He argues this is because the condition’s characteristic neurological damage disrupts the system in charge of generating endogenously-driven behaviour. The patient can respond to the environment’s exogenous affordances, but it cannot modulate the relevance of these affordances on the basis of internally generated processes.39 Contrasting normal agents with utilization-behaviour patients, Rietveld writes:

When the perceptual process of online tuning to relevant stimuli works smoothly, for example in skillful absorbed coping, even stimulus-driven action is to some extent an instance of endogeneously-driven behavior, because the sensory system ignores irrelevant stimuli and contributes to the perception of currently relevant stimuli. Our rapid unreflective responsiveness to affordances in absorbed coping probably has this relevance-tuned perception-action loop at its core. (Rietveld 2012)

Thus the capacity for Dreyfus-style skilful coping is based on the ability to ignore environmental features that are potentially significant, but that result irrelevant given the agent’s current goals and concerns. In other words, even the most modest cases of everyday skilful intentional action depend on the capacity to modulate attention, and link sensory and motor associations, in accordance with a concurrent intention. Rietveld calls this behaviour “rapid unreflective responsiveness”, probably intending ‘unreflective’ to mean ‘independent of calculation or deliberation’. But given the conception of ‘reflection’ as the executive inhibitory control of attention, the production of habitual, everyday skilful action, even in its minimal form as skilful coping, is reflective, insofar as it depends on the top-down modulation of attention.40

To move toward a conclusion, let us go back to utilization behaviour. Utilization behaviours are correct: the patients use objects skilfully, in accordance with automatic normativity

39 Similarly, L’Hermitte (1983) claims that “[t]he tactile, visuotactile and visual presentation of objects compels the patients to grasp and use them”, and stresses the point that behaviours are performed “without any internal motivation”.

40 Rietveld appeals to Freeman’s attractor theory (§2.2.1) for a non-representationalist account of how a neural system ignores stimuli that are irrelevant to the agent’s concurrent concerns. Freeman’s concept of “preference” is crucial in this explanation: “[p]reference provides an order parameter that shapes the attractor landscapes, making it easier to capture [...] desired stimuli by enlarging or deepening the basins of their attractors.” (Freeman 2000, 108). Proper discussion would require much lengthier treatment, but for now I suggest that the role of preference in Freeman’s attractor theory is broadly analogous to the role of intention as modulator of attention in Wu’s theory of action (the key role of both is to link relevant sensory stimuli to relevant motor behaviour). According to this, Rietveld’s view would seem analogous to the intellectualist view defended here, since intentional action even at the level of skilful coping would require top-down attentional modulation. (That said, Rietveld’s analysis of utilization behaviour has other, rather anti-intellectualist implications for the theory of agentive control. See §5.4.2 below.)
(e.g. they do not drop the coffee cup, but sip from it appropriately). So they display automatic control and normativity. They also thereby seem to display motor intentionality, in the sense that they count as alleviations of felt tensions, appropriate responses to affordances, or motor completions of a perceived environmental pattern. And they are clearly intention-independent: there is no endogenous goal that the patient is advancing—this is precisely why they are utilization behaviours: because they are not driven from inside, but rather exogenously originated. Because of this, although these behaviours are automatically controlled and motor-intentional, we should not call them ‘actions’—they are mere reactions to the environment’s affordances, exogenously-driven automatic responses.

But if we deny the name ‘action’ to these behaviours, why should we consider actions any other behaviour that is automatically controlled and intention-independent? There seems to be no clear difference between utilization behaviour, Wu’s fencer aunt, and Bratman’s spontaneous ball-catcher. They all seem to be equally exogenously generated. Mere automatic control may generate intelligent behaviour (like the expert fencer’s automatic response to her niece’s lunge), but it fails to generate intentional action if the behaviour is not an expression of endogenous significance, but mere response to an entirely disconnected affordance.  

So automatic control seems perhaps necessary, but certainly not sufficient for intentional action. For something to be called action, it has to be “endogenously driven”. It is probably very difficult to provide a general explanation of what it means for behaviour to be endogenously, as opposed to exogenously, driven. But it seems clear that some cases (like utilization behaviour, slips, and other such comportments) belong to the latter group. On the other hand, in a minimally complex environment, producing endogenous-driven behaviour implies the ability to focus on some environmental features and ignore others. This is a task impossible to perform without a minimal capacity for executive control of attention, i.e. the modulation of attention with a view to a higher-order endogenous goal. To the extent that this kind of top-down attention-control can be called ‘reflection’ (and I think it always should, since it requires the use of limited-resource executive functions (§3.2.2)), any behaviour worthy of the name ‘intentional action’—even of the most habitual, everyday, apparently mindless kind—requires reflective control.

---

41 Bratman and Wu’s cases are tricky because of the way they are framed. If the fencer’s response to her niece’s lunge is “playful”, as Wu describes it, it may count as endogenously-driven, in the sense that it partakes in the familial bond between them. Likewise, if I catch the ball that you throw to me, I may do so as a playful response to a playful gesture, or a defensive response against a sign of aggression. So it is hard to precisely draw the line in specific cases between what counts as purely exogenous and what is already endogenously-driven. But I hope the theoretical distinction, and the extreme cases of each, are sufficiently clear.
It follows from this analysis that distraction is the enemy of intentional action. I understand ‘distraction’ roughly as the capture of attention by affordances of the practical environment that are goal-irrelevant and detrimental to the goal-attainment. Utilization behaviour seems to be the extreme version of distraction is. Less radical versions of distraction are slips, procrastination, and mind wandering. These latter behaviours, if mild, do not destroy intentional action, but sufficiently potent episodes of any of them can lead to the disruption of goal-structured behaviour. The more long-term and complex the goal, the easier to disrupt it.

**The possibility of fringe actions**

As a final note, I would like to mention that this radical opposition between action and distraction, and the related conception of action as always hierarchically determined, seems to have some likely exceptions. I am thinking, for instance, about aesthetic experiences—where one freely allows for intuitive resonance with an object or a situation—, meditation—where one seeks to open the focus of attention to experience in general, without restricting its focus to what is goal-relevant—, or playfulness and improvisation—in which one purposefully includes task-irrelevant thoughts and behaviours into the action. These behaviours seem to count as actions, insofar as they are arguably goal-oriented (we have most aesthetic experiences when we intend to have them, we meditate when we intend to meditate, etc.), but they are peculiar in that a part of the action consists in manipulating attention itself: purposefully spreading it more broadly than it is when modulated by a standard intention. In a future occasion, I would like to study the relationship between ‘fringe actions’ like these and the general approach to action, automaticity, and control defended above.

For now, however, a pressing question remains open. I have argued that all intentional actions require reflection because action implies the modulation of attention in relation to a goal. But where do goals come from in the first place? How are they originated? Next Chapter tries to approach this question through the study of the cognitive phenomenon called ‘framing effects’.
Chapter 5: Automaticity and control in decision-making

The previous Chapter argued that even skilled and habitual actions are intention-dependent. This seems sufficiently solid ground to believe that all intentional actions must be coordinated top-down by an intention. However, this raises a further question that seems urgent to understand agency: where do intentions come from? And are we in control of this intention formation process?

In this Chapter, I will argue that we cannot conceive of ourselves as proper agents unless we are in some minimal control of our intention-formation process, and that such control is largely automatic. This entails a comeback of sorts for anti-intellectualism, which will be analyzed later on. Another upshot of this argument will be that individual agentive control is a matter of social and political concern, because the control we have over our intention-formation processes is significantly dependent on our practical environment, and this environment is shaped to a large extent by other agents, including public institutions. So attaining and maintaining individual agentive control turns out to be partly a matter of concern for public policy.

I will proceed as follows: after arguing that (§5.1) intention-formation control is a necessary component of even the basic forms of agency, and we can study intention-formation processes by focusing on decision-making processes, (§5.2) I introduce the phenomenon of framing effects to explain how our control over decision-making is fragile and rather limited. Given this, it is crucial for agentive control to find ways to shield our decision-making processes from the threats posed by framing effects and other cognitive biases. A study of the “debiasing” literature suggests that (§5.3) modifying the decision-making environment is much more effective than modifying the decision-maker when it comes to shielding control from framing effects. This analysis allows us to (§5.4) complete the discussion of agentive control with an analysis of the kind of control we have over intention-formation processes, its limits, and what we can do to enhance it.
§5.1. Intentions, decisions, and the problem of autonomy

5.1.1 Basic agency and the problem of autonomy

After introducing his situated-intellectualist position (of the top-down intention-structuring kind we have discussed in earlier chapters), Wu briefly mentions the problem of the origin of intentions:

The point of the current perspective is that control just is the role of intention in structuring a solution to the Many-Many Problem, full stop. The question of how the intention arises is an important one, but not one about agency in the basic sense. Rather, it pertains to whether the resulting action has higher-ordered properties such as whether it is free, rational or moral. (2013a)

Is the problem of intention formation irrelevant to basic agency? This, of course, depends on what we count as 'basic'. For Wu, what is basic about action is that we are able to structure it via intentions to solve Many-Many Problems, i.e. to build appropriate bridges between relevant sensory inputs and relevant behavioural outputs (cf. §4.5.1). But is that all there is to basic action? Consider the possibility that someone goes under deep hypnosis, of a kind that turns the agent fully cooperative with someone else’s intentions (Wilkinson 2013, 353).¹ Would we call the person under this influence an agent, even in the basic sense? Wu may insist that we should, given that the person can still effectively solve Many-Many Problems in accordance with the instructions given to her. But that is precisely the point: if this is all that is needed, then any machine complex enough to follow someone else’s instructions by selecting the relevant outputs for the relevant inputs (e.g. a smartphone) may qualify as an agent. Something else seems missing even for basic agency. Something like control over one’s own intentions.

This is indeed an intuition developed by the enactive approach to agency, according to which a basic agent is characterized by its autonomy: its ability to make sense of its environment in accordance with its own self-determined norms (Barandiarian et al. 2009, §2.1.2E). We would say that the hypnotized person is moving in accordance with an intention, but is merely being manipulated rather than acting, because although she controls her motions in accordance with an intention, she lacks the ability to control that intention herself—she is not self-governed.

¹ This is, of course, the folk myth about how hypnosis works. If it seems too unlikely, consider scopolamine, a drug known to generate transient global amnesia and submissive behaviour (Ardila & Moreno 1991). Scopolamine was popular in Colombia during the 80’s and 90’s to ‘hypnotize’ people into emptying their bank accounts and homes for robbers, all of which they do diligently without the ability to question or oppose their plan.
Given this, I hold that *autonomy*, understood broadly as requiring a significant level of control over the formation and endorsement of intentions that guide one’s behaviour, is a necessary condition for minimal agency.² If this is so, then our description of agentive control is not complete until we explore how, and to what extent, human agents are in control of their own intention-formation processes. Those are this Chapter’s main questions.

### 5.1.2. Intention formation and decision-making

I will approach the problem of intention formation via the problem of decision making, by considering intentions and decisions to be largely interchangeable. Recall that ‘intentions’, broadly construed (§2.2.3), are mental pro-attitudes (propositional or non-propositional, conceptual or non-conceptual) whose representational content includes a goal to be achieved by the agent, and imply a commitment on the agent’s part to achieve that goal. In that sense, an intention to X is very similar to a decision to perform X, in that they both involve the agent’s commitment to the represented object, and imply a resistance to the commitment’s being revised or modified (§2.2.3; cf. Bratman 1987).

To the extent that the concepts of ‘intention’ and ‘decision’ can be usefully run together, the psychological study of decision-making may shed light on the process of intention formation. It may at first sound like ‘decision-making’ is too reflective a term to capture the broader phenomenon of intention generation; but, as will become apparent in the next section, there is nothing necessarily reflective about decision-making, and even decisions we take to have made reflectively turn out to be shaped by automatic processes that may be significantly independent from reflection. Specifically, decision-making may be automatically susceptible to cognitive illusions analogous to perceptual illusions.

² The concept of autonomy cannot be treated here in its full complexity. I merely use it following the intuition that autonomous people (or communities) are those who govern themselves by establishing their own laws and objectives.
§5.2. Framing effects on decision-making: The limits of agentive control

5.2.1. Perceptual illusions and cognitive illusions

Have a look at Figure 5.1. Which of the two central circles has a greater diameter? It certainly appears like the one to the right does. But in case you are not already familiar with the image, you may have guessed from the subtitle that it is the wrong answer: measuring both circles reveals they have the same diameter. Something surprising about illusions like these is that even after reflectively unmasking their illusory nature their grip on intuition persists: we still see the right one as bigger, even though we know it is not.

This illusion, like many others, is all about context: we perceive the circle to the right as larger because, in comparison with the surrounding circles, it looks smaller, whereas the circles around the one to the right make it look larger. This turns out to be an epistemological failure on our part: we let our judgment about the central circles be influenced by irrelevant contextual features. But it is an understandable failure: contextual features automatically affect our cognitive processes in ways that are often helpful, and there is no way for us to shut off that effect in cases in which they are not.

![Figure 5.1: The Ebbinghaus illusion.](image)

Take a look at Fig. 5.2 now. By adding a slight environmental modification, it becomes much easier to arrive at the correct judgment. The rectangle helps us abstract the two circles from their confusing contexts, and compare them in a different context, no longer determined by the surrounding circles. Even in this new image, however, the illusion has some intuitive grip, and arriving at the correct judgment still requires a certain amount of measurement and
inference. But the inference is noticeably easier to make here than in the original version. In extended-mind terms (§2.1.2F), the rectangle provides cognitive scaffolding that extends our reflective capacities, thus making it possible for us to more effectively reason beyond the processing limits of what our inner capacities would allow (like pen and paper extend our math abilities, or a to-do list extends our practical skills).

Going back to Fig. 5.1, it can also be said that the perceptual illusion leads to a bias in our judgment: intuitive processes complete patterns that may or may not be relevant to the current reflective task, and may bias reflective deliberation toward producing a wrong answer. The rectangle in Fig. 5.2 thus works as a debiasing tool, which helps the agent in avoiding the biasing influence by offering a different perspective from which to interpret the situation.

I will argue in this section something similar happens in the realm of decision-making: like perceptual illusions can lead to judgment errors, (§5.2.2) cognitive illusions may lead to decision-making errors: just as in this perceptual illusion two identical circles look different because of varying (but irrelevant) contextual features, in the phenomenon of framing effects we have two identical choice-problems that look different because of varying (but irrelevant) contextual features. Further, (§5.2.3) framing effects seem to be generated in the intuitive processes that inform further reflection, and (§5.2.4) they seem to be widespread enough in practical decision-making realms that they become an actual threat to agentive autonomy.

3 In their seminal work on framing effects, Tversky and Kahneman themselves make the perceptual illusion analogy: “In the persistence of their appeal, framing effects resemble visual illusions more than computational errors” (1986, 260). The fact that framing effects are intuitive rather than computational errors is crucial for the argument I am building here. (Problems 1 and 2 below are taken from the same source.)
5.2.2. Two cases of cognitive illusions

Consider the following problems. What would you choose?

Problem 1
Assume yourself richer by $300 than you are today. You have to choose between
(A) a sure gain of $100
(B) 50% chance to gain $200 and 50% chance to gain nothing

Problem 2
Assume yourself richer by $500 than you are today. You have to choose between
(C) a sure loss of $100
(D) 50% chance to lose nothing and 50% chance to lose $200

In Problem 1, most people choose option A; in Problem 2 most people choose option D. But the two problems are essentially the same, since their two options are (extensionally) identical: they both ask for a choice between either $400 for sure or an even chance of $500 or $300. (A is identical to C; B is identical to D.) The difference between them is that, because the starting point varies (in Problem 1 it is $300, but in Problem 2 it is $500), the options in Problem 1 are both framed as gains, but the options in Problem 2 are framed as losses. Because the options are ultimately identical, shifting one’s preferences from one problem to the other implies a failure of rationality: we allow for an irrelevant contextual feature to alter our decision. This decision reversal breaks one of the core assumptions of rational choice theory, descriptive invariance, which states that a rational agent should make the same choice for the same problem even when presented under different descriptions. If the options given to the agent are found, upon reflection, to be the same, then the agent’s choice should be the same throughout the problem’s diverse descriptions. A number of studies show that we tend to break the rule of invariance due to differences in the problem’s framing.4

Now consider the following story from economist Richard Thaler. As a young lecturer, Thaler had a whole class of students complain that his first exam was too difficult: they deemed the average (72 out of 100) to be too low. To remedy the problem without lowering the difficulty level, Thaler decided he would make the total possible points 137 instead of 100. The

---

4 For the classical demonstration that invariance does not hold as a descriptive rule of human decision-making, see Tversky & Kahneman (1981). For recent discussion, see Stanovich (2010, 22–36).
next exam’s average score was 96, and students went from angry to happy. But an average of 96 out of 137 is equivalent to 70%, so the average was practically the same in both exams. So why were the students happy? As Thaler explains, a 96 sounds positive to someone habituated to a maximum of 100, and very few students seemed to bother converting the grade into a percentage—an effortful calculation to do in your head. This is, again, a cognitive illusion: 72 out of 100 and 96 out of 137 are essentially the same average score, so if the students were right to be angry the first time, they should also be the second time.

Concerned with seeming manipulative, Thaler adds:

Lest you think I was somehow deceiving the students, in subsequent years I included this statement, printed in bold type, in my course syllabus: “Exams will have a total of 137 points rather than the usual 100. This scoring system has no effect on the grade you get in the course, but it seems to make you happier.” And indeed, after I made that change, I never got a complaint that my exams were too hard. (2015, 4)

Regardless of whether he was manipulating his students or not (more on this below), it seems clear that there is a failure of rationality here: students shifted their preferences with respect to an identical problem due to an irrelevant contextual feature (the original average looks like a loss; the subsequent one, like a gain).

5.2.3. The cognitive mechanisms behind framing effects

Framing and valuing

The two former cases reveal the existence of framing effects, i.e. shifts in preference due to differences in how the problem is framed. This effect has been confirmed by many studies since Tversky and Kahneman’s original presentations. Essentially, what framing effects tell us about the way we make decisions is that we choose not from an objective standpoint (assessing options merely by the amount of absolute value each provides), but in comparison with an initial

5 Strictly speaking, ‘framing’ refers to a problem’s precise wording. More loosely, it may also refer to other contextual features of the problem’s presentation that may influence the problem’s interpretation. In the literature, the word ‘framing’ is also used in another sense, to refer to the agent’s cognitive processes of problem interpretation. To avoid ambiguity I use ‘framing’ only for the problem’s presentation, and use ‘interpretation’ or ‘sense-making’ to refer to the agent’s cognitive processing and interpretation of the problem. (For discussion of definitional issues, see Kühberger (1998); Levin et al. (1998); Gold & List (2004); and Maule & Villejoubert (2007).)

6 A meta-analysis by Kühberger (1998) suggests that the framing effect is reliable, although effect sizes vary depending on design variables, particularly the manipulation of risk (as a probability that the agent’s choice yields losses vs. gains, or as a risk implicit in the current situation [e.g. the chances of suffering a particular disease] that the agent may or may not actively confront), and the participant’s task (i.e. making a choice vs. making a judgment).
subjective state, which we take as the status quo, relative to which we interpret each option as a loss or a gain. Crucially, we value the same option differently when framed as a loss than when framed as a gain: avoiding a loss of a certain magnitude seems more choiceworthy to us than pursuing a gain of the same magnitude. (This explains why the students are angry when their grade average appears to be a loss, but no longer care when the same average appears to be a gain.) Moreover, in situations of risky decision-making, the imbalance in our attitudes toward gains and losses leads to more risk-averse choices when choosing between gains, and more risk-seeking choices when choosing between losses. (This explains why most people choose the certain option in Problem 1 [where options are framed as gains] and the risky option in Problem 2 [where options are framed as losses], despite both problems being identical.) Such difference between the way we value losses and gains is often called loss aversion. The famous “Asian disease problem” provides another example of loss-aversive behaviour:

Problem 3
Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
- If Program A is adopted, 200 people will be saved.
- If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

---

7 Loss aversion is illustrated by the shape of the value function. First, value is described in terms of gains and losses relative to the initial state. Second, the function is not linear, but concave for gains and convex for losses; so the value function is steeper for losses than for gains (thus, for any amount \( x \) the value of gaining \( x \) will be smaller than the disvalue of losing \(-x\)). Thus, people are usually risk-averse for gains and risk-seeking for losses. (See Fig. 5.3).

---

Figure 5.3: The value function proposed by Tversky & Kahneman (1981)
Most people facing this version of the problem (where options are framed as gains) choose the first, risk-averse option. However, people facing equivalent options expressed in terms of losses behave differently:

Problem 4:
- If Program C is adopted 400 people will die.
- If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

Here most go for the latter, risk-seeking option. Moreover, even participants who see both versions of the problem experience a tension between their wishes to remain risk-averse in Problem 3 and risk-seeking in Problem 4, on the one hand, and to be consistent on the other. In this way, framing effects are analogous to perceptual illusions: even when we reflectively realize their illusory character, the irrational answers retain their intuitive appeal.

**A dual-process model of framing effects**

It is often claimed that problems of irrationality in judgment and decision-making are due to excessively relying on type-1 processing, when we should be thinking things through reflectively. If this is so, then getting people to reflect, or reflect more intensely, should contribute to avoiding irrationality by overriding the defective intuitive outcome and substituting it for an appropriate reflective one. This may be true in many cases, but when it comes to framing effects things seem different. I will examine evidence that suggests increasing reflection does not shield agents from framing effects below (§5.3.2). For now I introduce a dual-process model of framing effects that will be handy later on.

Decision-making seems to occur in two stages: we first produce a preliminary interpretation of the decision situation, narrowing its multiple features down to the ones that seem relevant; then we compare the options we have interpreted as relevant, and select the one with highest apparent value. The first stage seems to be mostly intuitive: it consists in the typical associative pattern-completing, tension-alleviating process (§3.4). Through automatic-intuitive processing, some features become salient and accessible for further processing; some of the

9 Tversky & Kahneman (1986, 260). Notice the similarity with Gendler’s description of alief-belief discrepancy (§2.2.1).
10 See Shafir & LeBoeuf (2002, 499–500) for a review of research endorsing this view. A similar claim is endorsed by Stanovich (2004), and in philosophy by what could be called the Socratic tradition (§3.3.3).
relevant options appear as gains and others as losses, depending on the problem’s specific framing; and some other features are inhibited and rendered inaccessible for further processing. Once the general practical situation is intuitively interpreted, the second stage can take place, in which the agent evaluates the options and chooses one. In rapid decision-making, this stage may be entirely intuitive; in situations requiring some calculation or hypothetical thinking, reflection may be recruited to compare the alternatives’ relative value.

Back to the framing effects–perceptual illusions analogy. Upon seeing the Ebbinghaus illusion, if someone judges the right circle to be larger than the left circle, this shows that the surrounding circles have influenced her initial, automatic-associative interpretation of the situation. Similarly, in framing effects the cognitive illusion takes place because some (objectively irrelevant) features of the situation affect the first-phase intuitive process of interpretation, particularly the attribution of relevance. If this is so, framing effects occur not because we produce intuitive judgments or decisions when we should be reflecting, but rather because the first phase of decision-making is susceptible to cognitive illusions (i.e., to irrelevant contextual aspects seeming relevant, or relevant aspects seeming irrelevant), regardless of whether we reflect or not afterward. A consequence of this is that a decision can be made reflectively and effortfully, and still fall into a cognitive illusion: I may reflect for a long time about which program is better for facing the Asian disease without realizing, or even considering, that my reflection is biased because the programs are formulated in terms of lives lost.11

5.2.4. Framing effects and the limits of autonomy

Framing effects can plague people’s everyday decisions, both trivial and transcendental. This may lead us to lose some degree of control over the decision-making process, so that, given sufficient framing effects, our autonomy may be in risk. In this subsection I mention cases in which framing effects alter everyday decisions, showing that framing effects can lead to losses of self-control, and make us susceptible to manipulation by people who control the way options are framed. If this is so, framing effects present an important challenge to agentive control.

---

11 For this two-stage view of decision-making, see Tversky & Kahneman (1986); Evans (2008, 263–264). Evans claims that in decision-making “[p]eople are using their reflective minds […], but the problem leading to framing effects is that the information, which appears to them to be relevant, has been predetermined by intuitive and unconscious processes.” So “just because the reflective mind is involved in a decision, it does not mean that it is in charge” (2010a 89–90).
Framing effects and decision myopia

The structure of decision-making depicted in framing effects and loss aversion reveals that we can often suffer from a sort of decision myopia: the tendency to focus excessively on avoiding small, proximate losses may lead us to lose sight of larger, more distant gains. This phenomenon has several repercussions on the structures of our lives: we tend to let opportunities pass us by; we are in general bad at losing; we are bad cooperators in cases in which we would all be better off cooperating; and we are often bad at resisting temptations.

(A) We let opportunities pass us by

Given our loss-averse disposition, we tend to act more promptly when options are framed as averting losses than when they are framed as averting gains. Conversely, we may delay or neglect acting when options are framed as procuring gains, even if acting would yield great benefits in the long run. Consider this in the context of medical decision-making.

It has been shown that, among women who were encouraged to perform breast self-examination [BSE], those exposed to positively framed sentences were less likely to engage in BSE than those exposed to negatively framed sentences.\(^{12}\) This suggests that highlighting the benefits of a health measure is less effective in promoting it than highlighting the potential losses of not doing so. So it is likely that there are patients who could have caught the onset of breast cancer earlier on, thus making it more treatable, but did not because the BSE was framed positively (e.g. BSE increases longevity) rather than negatively (as, say not performing BSE decreases longevity). They lost the opportunity due to framing.

Lost opportunities can be due to which option is presented as the default. Recall that great differences in organ donation rates depend on whether a given country has an opt-in or an opt-out system.\(^{13}\) Investigating the influence of defaults on medical decisions, Narula and colleagues (2013) studied the case of colorectal cancer treatment. Researchers tested two different interventions for a group of 60-70 year-old patients whose last colonoscopy had taken place more than 10 years ago. In the opt-out intervention, patients received a letter explaining it was time for another screening, along with a date and time for an automatically scheduled colonoscopy (which they could change or cancel). In the opt-in intervention, patients received the explanatory letter with instructions to call and schedule an appointment themselves. 85% of patients in the opt-in condition showed up to their appointments, but only 62.5% in the opt-out condition did. So it can

---


\(^{13}\) Johnson & Goldstein (2003); see discussion of framing effects in the Introduction above.
be assumed that some high-risk patients have missed the opportunity of early detection due to differences in framing—particularly, due to which option was established as the default.\(^{14}\)

The influence of defaults is of course not restricted to medical decisions. In a widely discussed case, the inscription rates of employees in 401(k) pension plans in the U.S. vary considerably depending on whether the company has an opt-in or an opt-out system.\(^{15}\) Given the great difference that starting pension-plan contributions early makes in retirement leisure, it can be said that many employees of companies with opt-in systems may have lost the opportunity of a much more leisurely retirement partly because of the framing of options.

**B) We are bad losers**

Loss aversion makes us bad losers. According to rational choice theory, in making a decision we should not take sunk costs (i.e. costs that we incurred in the past and are irrecoverable) into consideration, but we should focus solely on future costs. But we do tend to consider sunk costs in decision-making: once we buy a theatre ticket we are more likely to go, even if we would rather not (because we already paid); if we bought two products but can keep only one, we tend to keep the more expensive one even if we would rather have the other one for other reasons; we tend to not sell an underperforming stock in time hoping that it will start doing better, which often leads to incurring further losses. This is the *sunk cost effect*, whose general pattern is that we tend to persevere in a losing endeavour. This can be explained by reference to our tendency to be more risk-seeking for losses than for gains. Because we do not like feeling like we have wasted resources, we refuse to accept a loss and persevere when it would be wiser to quit, and ‘throw good money after bad’.

So loss aversion generates a framing effect that makes us bad losers: we are more likely to make a losing bet if the bet is framed as a continuation of a past project. Sports teams tend to keep famous athletes past their prime for much longer than they should: these athletes demand a lot of money and no longer produce the expected results, so teams could find more and better players for a smaller cost, but they often fail to disregard sunk costs (like the money already paid for them). Countries fight hopeless wars till the bitter end. People persevere in harmful long-term relationships and unfulfilling career projects even though they would be better off quitting.\(^{16}\)

Thus, sunk cost effects are framing effects, because they make us less able to quit when quitting is seen as giving up, even though we would quit if we could see it as simply cutting costs. Therein lies the cognitive illusion, and another way in which framing effects can limit the

---

\(^{14}\) For a general discussion of framing effects on health-related decisions, see Rothman & Updegraff (2010).


\(^{16}\) On sunk costs, see Arkes & Blumer (1985); Zamir (2012, 839–840; 2014, 29–ff.).
control we have over our decision-making process: our preferences shift by irrelevant features of the practical environment.

(C) We are bad cooperators

Loss aversion and status quo bias also have automatic influences on our decisions whether to cooperate or not with others. Researchers have studied the influences of framing in *public goods dilemmas*, strategic interactions in which each agent must choose whether (and how much) to cooperate with others in providing or maintaining a public good (i.e. a benefit from whose use others cannot be excluded), where cooperating is at odds with maximizing individual benefit. The same cooperative strategy can be framed either as foregoing a gain or as incurring a loss. Evidence suggests that participants in these dilemmas tend to cooperate less when cooperating is framed as a loss than when it is framed as foregoing a gain. In other words, situations in which cooperation is framed as *giving* lead to less cooperation than those in which it is framed as *not taking* (although the sizes of these framing effects vary in different contexts).17 Even naming a game “Wall Street Game” (or “Stock Market Game”) can yield significantly lower cooperation rates than when the same game is named “Community Game”.18

These findings may help explain why deviance gains the status of social norm in some societies. When waiting in line at the bus station, following transit rules while driving, or refraining from corrupt practices, are seen as a loss (of status or of possible benefits), or when not cooperating is seen as the status quo, people will tend to engage less in cooperation than when such behaviours are seen as foregoing possible gains, or when cooperating is seen as the norm.

Framing effects have been used to explain phenomena at the much greater scale of international negotiations. To mention one case, framing effects based on loss aversion and status-quo bias make a successful climate change negotiation less likely.19 The gains of taking climate action (about which there still is much uncertainty, given their considerable temporal distance) are assessed against the losses of reducing emissions (which are much closer in time, thus more palpable), so loss aversion makes the latter loom larger, and status-quo bias makes people less willing to incur losses to current living standards. Moreover, in international climate negotiations each party is likely to frame

17 See e.g. Brewer & Kramer (1986); Fleischman (1988); Sell et al. (2002); Goerg & Walkowitz (2010). The effect is modulated by many features, among which are the game structure’s particularities (e.g. repeated vs. one-shot); the tendency to act as conditional altruists (i.e. to cooperate whenever the others are doing so too); and the cultural traits of participants. This makes this kind of framing effect more context-dependent and less reliable than others (particularly the ones closer in form to the Asian disease problem [see Kühlberger (1998) and Cubitt et al. (2011) for discussion]). Nevertheless, the visible trend is that loss frames tend to generate less cooperation than non-gain frames.

18 See Liberman et al. (2004); Ellingsen et al. (2012).

their own sacrifices as losses and other parties’ sacrifices as gains—a situation that would make each party averse to making concessions, thus making a stalemate more likely.

These considerations lead to the conclusion that framing effects make us bad cooperators: they make us less likely to cooperate in situations in which all parties would be better off by cooperating. In a word, framing effects can make collective action problems harder to solve. This is yet another way in which the cognitive illusions of framing effects pose a threat to autonomy: loss aversion and status-quo bias can lead us to making decisions that make us worse off, and that we may not endorse if we could see the practical situation from a debiased perspective.

(D) We lack self-control

Framing effects also occur in a temporal register. We pursue smaller gains that are closer to the present in a way that makes us disregard more distant, yet greater, gains; and negatively, we avoid smaller losses that are closer to the present in a way that makes us disregard more distant, yet greater, losses. This present bias can be seen as a status-quo bias applied to time: the options closer to the present loom larger than those further away.

Present bias generates framing effects in the sense that the temporal distance of options—a feature of the practical situation that could be considered irrelevant—makes us shift our preferences. Thus, today I judge that getting up early in the morning tomorrow (say, to go for a run) is better than sleeping in, and thus decide to wake up early tomorrow; but when the time to wake up comes, the benefit of sleeping in seems much greater than the benefit of getting up (or inversely, the costs of getting up now are much larger than the costs of not running today); so I turn off the alarm. Later in the day, I come to regret not having stuck to my original plan, because it becomes clear again that I do consider the benefit of running greater than that of sleeping in. Repeat this preference shift sufficient times, and we have a problem of self-control.

Just like previous cases, this phenomenon seems to depend on a cognitive illusion: we have an excessive aversion to satisfaction delays, which makes us discount a more distant satisfaction excessively when compared with a more proximate satisfaction when it is very close to the present. The fact that we attribute more value to the distant satisfaction is clear when both satisfactions are far into the future, or when the smaller-sooner satisfaction is already in the past. But our value function seems to be warped with respect to the satisfactions that are very close in the future: thus when a smaller-sooner satisfaction looms near enough, it suddenly becomes much more desirable than when it was further away. This phenomenon, called hyperbolic discounting,\(^{20}\) leads us to overvalue smaller

\(^{20}\) See Ainslie (2001). Crucially, the discounting function is not exponential (as rational choice theory would have it) but hyperbolic. The hyperbolic form is what explains the preference reversals that constitute self-control problems.
immediate gains to greater ones that are somewhat more distant (e.g. we tend to overvalue the smaller immediate rewards of spending all our extra money, watching t.v., or eating an extra sweet treat, to the longer delayed rewards of saving, being fit, or having a healthier diet); or it leads us to postponing smaller immediate sacrifices which in turn pile up into greater, more distant ones (e.g. we tend to put off grading a few papers each day, so we end up having to sit down and grading them all when the deadline approaches).

Hyperbolic discounting seems to be based on a cognitive illusion that distorts the value of options (it makes greater gains appear smaller, and smaller gains appear greater) when one of them is closer in time than the other. Such cognitive illusion generates framing effects—i.e. preference reversals when facing identical problems due to differences in temporal framing (see Fig. 5.4)—which ultimately lead us to losing some amount of control over our decision-making processes. Well-known and widely discussed problems of self-control, like procrastination and weakness of will, are among the losses of autonomy that can be explained by reference to the cognitive illusions and framing effects related to present bias and hyperbolic discounting.21

Figure 5.4: Hyperbolic discount curves (by Ainslie [2005]) of a smaller, sooner reward (e.g. sleeping in) and a larger, later reward (e.g. getting up early and running). The smaller reward appears larger as it becomes closer to the present; but the larger reward’s superiority becomes evident again later on. This explains the ensuing regret if the smaller reward was chosen.

Decision myopia can lead to manipulation

In sum, framing effects make us myopic decision-makers, who are likely to miss larger, longer-term benefits because we are focused on avoiding smaller, shorter-term costs. This myopic behaviour, which we can engage in naturally due to structural features of our intuitive practical cognition (like loss aversion and hyperbolic discounting), can also be intentionally used by

---

people in charge of structuring the framing of options (what Sunstein and Thaler (2008) call “choice architects”) to steer our decision-making processes in one way or another.

In fact, Thaler himself confesses to have done this to his own students: by altering the way information was presented, students went from angry and dissatisfied to happy or unconcerned (§5.2.2). Does this not count as a case of manipulation? And if so, is it a morally or politically problematic to engage in such manipulative practices? In this section I (E) discuss some cases of how choice architects can use decision myopia to manipulate agents’ choices, and (F) introduce some philosophical problems about how to interpret and evaluate manipulation.

(E) A wealth of opportunities for pulling people’s strings
Choice architects can use each of the cognitive biases discussed earlier to steer people’s decisions in a particular direction. They can use loss aversion, e.g., to frame medical messages in a certain way to increase preventive behaviour, or use the status-quo bias to determine opt-in or opt-out rates for a given program by manipulating which option appears as default (think about the organ donation or 401(k) examples discussed above). These seem benign and justifiable interventions, aimed at the public interest. But the same tricks can be used, and have indeed been used for a long time by the advertising and marketing industry to try and manipulate consumers’ choices. Some often-cited, and quite ubiquitous, examples are the quasi-framing effects achieved by “odd-pricing” a product (e.g. at $9.95 or $9.99 rather than $10),22 or the use of phrases like “For a limited time only”, “Sale”, and “only 2 seats left at this price” to frame buying as avoiding a potential loss (missing out on a good deal), thus exploiting loss-aversion.23 With a more systematic understanding of bias, marketing becomes more powerful. Recently, a marketing firm was asked to help a News UK call centre. The Times and Sunday Times newspapers had recently raised subscription prices, and the rate of customers calling to cancel subscriptions hiked. So marketing specialists taught call-centre employees to frame options according to behavioural insights while talking to customers. This included framing cancellation as a loss (“I wouldn’t want you to miss out on this special package”) or retaining membership as the status quo (“Many people like you have chosen this package”). According to the firm, when the call-centre employees purposefully used cognitive

---

22 Macé (2012); Choi et al. (2014).
biases in the conversation, the call was three times more likely to be successful, not only in avoiding cancelation, but even in persuading customers to upgrade their subscriptions.\footnote{The firm in question is Ogilvy & Mather; the numbers are reported by an Ogilvy employee (cf. “The Maddest Men of All”, a \textit{Freakonomics} podcast episode available at \url{http://freakonomics.com/2015/02/26/the-maddest-men-of-all-full-transcript/}).}

These more clearly manipulative strategies are not available only to the private sector, of course. Public-sector entities can steer decisions in many small ways that may or may not amount to manipulation. Here is a list of strategies that choice architects may employ to steer people’s decisions in a particular direction:

(a) Choice architects might choose a graphic health warning, on the theory that an emotional, even visceral presentation might have significant effects.

(b) […] They] might be alert to framing effects and present information accordingly.

(c) They might be aware that a statement that a product is “90 percent fat-free” has a different impact from a statement that a product is “10 percent fat,” and they might choose the frame that has the desired effect.

(d) They might make a strategic decision about how to present social norms, knowing that the right presentation – for example, emphasizing behavior within the local community – could have a large impact on people’s behavior.

(e) They might decide to list options – in a cafeteria or on a form – so as to make it more likely that people will make certain choices. (Sunstein Forthcoming, §119)

Which among these interventions should we count as manipulative? And which ones should we consider morally problematic?

\textit{(F) Manipulation and autonomy: Philosophical problems}\footnote{This subsection is especially indebted to Wilkinson’s (2013) analysis of the concept of manipulation.}

A difficulty when analyzing manipulation is that widely different things may count as manipulative. Lying to someone to generate a certain response is usually taken to be manipulation; but displaying a certain product more attractively than others (with flashy colours or signs, say) could also count as manipulation under some lights. Making sad faces to your partner or friends to influence can count as emotional manipulation in some contexts; altering the default rules of a system to increase enrolment in a service can be judged manipulative as well.

What makes a behaviour manipulative? We can start by distinguishing manipulation from coercion. According to Joseph Raz (1986, 377-378), “[m]anipulation, unlike coercion,
does not interfere with a person’s options. Instead it perverts the way that person reaches decisions, forms preferences or adopts goals”. Manipulation perverts or distorts decision-making, Raz holds, by interfering with the way the person forms preferences or adopts goals. This resonates with the above discussion of framing effects. A choice architect (which could be a public or private institution, a friend or a colleague, or even oneself) may use framing effects to manipulate choice by intentionally placing a cognitive illusion in an agent’s practical environment, with the expectation that such illusion would lead the agent to form a certain preference arrangement, or set a certain goal, rather than another.

Raz considers such decision-making distortions morally problematic to the extent that they constitute “an invasion of autonomy”, a problem that goes beyond the manipulation’s consequences: even if the manipulative intervention makes the agent better off in some respect, intentionally undermining her control over decision-making remains questionable because it implies treating her like an object rather than a person, like a means rather than an end in herself. But exceptions must be acknowledged: manipulation and even coercion are morally permissible when intended to prevent harm to others (as in stopping someone from crossing a street when a car comes straight toward them, or hiding weapons from a friend who is in the middle of a fit of rage—to use Plato’s example).

Exceptions can be found elsewhere. Lying is often considered a paradigmatic case of manipulation, but even some cases of lying seem ethically unproblematic, like when your friends manipulate you to throw you a surprise birthday party. We often manipulate ourselves in subtle ways by arranging our environments so as to make some options more salient than others, by e.g. hiding the junk food deep in the counter while leaving the fruits and vegetables in visible, accessible locations. These processes of environmental manipulation offload the cognitive processes required to exercise self-control, thus effectively solving some cases of decision myopia. Moreover, many of us may appreciate our governments doing similar things, like setting up automatic enrolments for uncontroversially beneficial plans, at least if the opt-out processes are simple enough. But what if default rules are manipulated and opt-out processes are made more difficult? What if a government uses loss aversion to gain popular support for controversial policies, to halt reforms, or to benefit the ruling party’s permanence in power?

It seems clear that some intentional alterations of choice architecture are manipulative in a way that poses no moral problems: we do this to ourselves to avoid self-control problems, and (would) consent to our governments doing the same in our benefit. Yet it also seems clear that

---

26 On this see also Darwall (2006, 269); White (2008).
other such alterations would be manipulative in morally problematic ways. And many others are in between these extreme cases. Moreover, some ways of altering the choice architecture seem autonomy undermining, but others seem autonomy enhancing. Can clear distinctions be made between acceptable and unacceptable choice-architecture modifications? The next sections offer elements toward an answer. (I come back to this issue at §5.4.4.)

**Decision myopia, manipulation, and autonomy**

Loss aversion, status quo bias, and present bias are constitutive aspects of our ‘practical perception’ capacities, i.e. our capacities for making automatic and intuitive sense of the practical world. They seem to have developed early on in our evolutionary history, and probably made sense in much of our species’ history, but they become less and less useful as our environments are altered by the technologies and institutions of modern life. In modern, late-capitalist contexts, such practical-perception structures threaten to undermine our autonomy by presenting relevant features as irrelevant, and irrelevant features as relevant, in a way that can make us let opportunities pass by, unable to accept loss, unable to cooperate, and susceptible to losing self-control—all of which are different ways in which we can undermine our own autonomy by losing control over our decision-making processes. Moreover, these structures of practical cognition make us susceptible to being manipulated by choice architects, i.e. institutions or individuals (including ourselves) who get to determine the way our options are framed.

If this is so, what should we do to protect or augment the cognitive capacities that allow us to control our decision-making processes? In the following section I introduce the notion of *debiasing techniques*—the techniques that have been proposed to counter cognitive biases, thus shielding autonomy from them—, and analyze which debiasing techniques result more useful in debiasing framing effects in particular. This will shed light on the nature of the decision-making process and the kind of control we can have over it.

---

28 See Zamir (2012, 840–841; 2014, 42-46); Heath (2014); Soll et al. (Forthcoming).
§5.3. Debiasing framing effects

This section (§5.3.1) presents the main types of debiasing techniques, and explores which of these are more successful in debiasing framing effects. I argue, on the one hand, that (§5.3.2) when it comes to debiasing framing effects, debiasing techniques that seek to modify the decision-maker (by increasing skill or intensifying reflection) have proven rather unsuccessful in this task. But on the other hand, (§5.3.3) modifying the environment seems a much more promising approach. This rather surprising feature of framing effects provides (§5.3.4) a chance to make some philosophical remarks on the nature of framing effects and decision-making.29

5.3.1. Debiasing techniques

Conceiving ‘biases’ broadly as deviations from an objective decision-making standard (like e.g. rules of rationality or probability), debiasing techniques are understood to be methods aimed at cancelling the normatively negative effects of cognitive biases. In their recent review of the debiasing literature, Soll and colleagues distinguish two main approaches to debiasing:30

• “Modify the decision-maker”, or decision-maker debiasing: The occurrence of a bias may be due to the agent’s lack of skill, knowledge, or disposition to reflect in a particular situation. Hence it may be useful to ask agents to think about the choice problem more thoroughly, or to provide them with relevant training. Thus there are two distinguishable approaches:
  • Synchronic decision-maker debiasing instructs, encourages, or provides opportunities for the agent to engage in reflection at the specific time of making the decision.
  • Diachronic decision-maker debiasing provides the agent with training or skills relevant for making the decision in question more proficiently. This includes teaching numeracy skills, like the handling of probability information, or useful rules of thumb (e.g. “Think the contrary”) and effective checklists.31
• “Modify the environment”, or environmental debiasing: The negative effects of bias can also be due to a mismatch between the choice environment and the agent’s cognitive capacities, whenever environmental features contribute to the appearance of cognitive illusions. In these situations it is helpful to alter the choice environment so that the agent is able more

---

29 This section’s argument is restricted to debiasing framing effects. The conclusions would surely be different for other biases, e.g. hindsight bias or overconfidence (on which see Fischhoff 1982). For a related argument concerning present bias, hyperbolic discounting, and procrastination, see Heath & Anderson (2010). For general discussions of debiasing tools, see Larrick (2004) and Soll et al. (Forthcoming).

30 See Soll et al. (Forthcoming). For alternative taxonomies of debiasing techniques, see Larrick (2004, 316–317) and Zamir (2014, 38–42).

31 On the debiasing power of rules of thumb and checklists, see Larrick (2004; 2009).
fruitfully to engage with it. Such ‘environmental debiasing’ techniques can be divided into two main categories:

- **Information design** alters the way information is displayed, so that the agent processes it more effectively (by attributing salience to relevant features and not to irrelevant features).

- **Rebiasing** seeks to neutralize the negative effects of biases by countering the latter with other biases, so that the option that tends to seem more appealing coincides with the one that leads to better outcomes for the agent.

To illustrate this distinction, recall the Asian disease problem presented above. One may attempt to eliminate the framing-generated biases by giving participants a crash course in probability, so that they were able better to compare the two alternative programmes (which is a *diachronic decision-maker debiasing* technique). One may instead give them an explicit instruction to make sure to reflect about the problem, and even ask them to write down a justification for the choice they made (which are *synchronic decision-maker debiasing* techniques). Yet another way to aim for debiasing would be to provide participants with visual representations of each programme (in diagrams or pie charts), which may allow for them to assess the programmes independently from the gain- or loss-framing (which would constitute an *environmental debiasing* approach of the *information design* kind). Finally, one may try and cancel bias by setting one of the alternatives as a default (e.g. by adding something like “Program A will be chosen if nothing else is done; you can check the box below if you would rather choose Program B.”), thus using another cognitive bias to steer agents toward one of the alternatives rather than the other (which would be an *environmental debiasing* approach of the *rebiasing* kind). Agents and institutions can, of course, combine any of these debiasing approaches to improve decision-making in a given situation.

Now, to return to the main topic, which of the two debiasing approaches functions better in the particular case of framing effects?
5.3.2. How not to debias framing effects

The mind has its illusions as the sense of sight; and in the same manner as feeling corrects the latter, reflection and calculation correct the former.

—Pierre-Simon de Laplace

Because experts are often able to make much faster and more accurate decisions than non-experts, and because in many situations people tend to solve problems more correctly if they think harder about it, it makes sense to think that increasing expertise or increasing reflective effort are good debiasing candidates for framing effects. However, and rather surprisingly, evidence tends to suggest these usual candidates do not help in the case of framing effects. So Laplace’s statement seems to be false, at least regarding the illusions of framing effects.

**Expertise does not reduce framing effects**

Evidence suggests that experts can be deceived by cognitive illusions just as much as (or sometimes even more than) novices. Here are some examples: Patients, graduate students, and physicians were found to be susceptible to framing effects in choosing between two alternative cancer therapies. Professional intelligence analysts are more susceptible to framing effects regarding the Asian disease problem than college students. In a recent study worth some detailed consideration, the World Bank (2015, 182–183) sought to assess whether their own staff’s decision-making was susceptible to framing effects in their area of expertise: development economics. They presented different groups of staff with two identical problems, one framed as a decision about which of two skin creams was more effective in reducing a rash, and the other as a question of whether minimum-wage laws led to an increase in economic growth or unemployment. They found that World Bank staff gave more accurate responses in the skin-cream frame than in the minimum-wage frame, despite the structure of both cases being identical. They turned out to be particularly inaccurate when the data conflicted with their own

---

32 As quoted by Larrick (2004, 316).
33 Tversky and Kahneman (1986, 257) hold that the intuitive stage of decision-making is shaped partly by the person’s habits, skills, and preferences, and thus that developing expertise is a promising path toward making debiased decisions (see Shafir & LeBoeuf (2002, 502) for other similar positions).
34 McNeil et al. (1982).
35 Reyna et al. (2014). Cf. also Neale & Northcraft 1986; Dawes et al. (1989) [concluded that actuarial decision making is often far superior to expert, or clinical, judgment]; Redelmeier & Shafir (1995).
36 Participants had to report their level of support for the claims “Incomes should be made more equal” and “We need larger income differences as incentives for individual effort”. This study replicated the methodology used by Kahan et al. (2013) to report a similar finding.
ideological commitment (i.e. minimum-wage-law endorsers were most inaccurate when the problem showed minimum wage laws increased unemployment in that case). Here expertise led to an *increase* rather than a decrease of the framing effect, perhaps because it creates opportunities for ‘motivated reasoning’, i.e. reasoning subordinated to one’s goals or commitments. As the report comments:

One might be tempted to add that this occurred even though many World Bank staff members are highly trained experts on poverty, but in reality this occurred *because* World Bank staff members are highly trained on that topic. Faced with a demanding calculation, they interpreted new data in a manner consistent with their prior views, about which they felt confident. (2015, 18–19; my Italics)

This is a striking result, which suggests that although experts are often able to produce much more accurate judgments than novices, the former may be even more susceptible than novices to framing effects in cases in which the evidence goes against their broader theoretical or ideological commitments.

In a similar situation, Schwitzgebel and Cushman (2012) found that experts in moral philosophy were susceptible to the way different versions of the Trolley problem were presented. Specifically, their judgments about two Trolley-Problem scenarios (widely known as ‘lever’ and ‘footbridge’) varied significantly depending on which of the two scenarios was presented first.37

Thus, available evidence seems to show that diachronic decision-maker debiasing, in the form of expertise, does not contribute to debiasing framing effects. This suggests that framing effects are not caused by a lack of familiarity or involvement with the problems, since even those expert in the field are susceptible to them (cf. Shafir & LeBoeuf 2002, 502). Familiarity may even contribute to increasing some cognitive biases, insofar as overpracticed associations triggered by contextual cues may lead to narrow thinking in familiar terms that seeks to confirm unquestioned assumptions, and to disregarding atypical but important situational features (Chi 2006, 26–27).38

---

37 The specific wording of each case was not varied, so this would not count as a framing effect in the narrowest version of the latter term (where ‘framing’ is taken to mean ‘wording’). But in a broader conception of framing, taken to include some non-literal features of the decision situation, the susceptibility of expert philosophers’ judgments to an *ordering effect* also counts as a framing effect. The authors have also provided evidence regarding the susceptibility of philosophers to standard framing effects (see below).

38 This is not to say that education and expertise do not contribute in debiasing in general; they often do. The argument is restricted to debiasing framing effects in particular. (See Larrick (2004) and Soll et al. (Forthcoming) for broader discussions of the benefits and limits of diachronic decision-maker approaches like education and expertise.)
Increasing reflection does not reduce framing effects

Since framing effects seem to occur in response to shallow features of the situation (like wording or ordering), it has often been suggested they could be reduced if people paid more attention to the situation’s ‘deeper’ features, which should be achieved simply by thinking more carefully. This hypothesis received initial empirical support by studies showing a correlation between higher levels of Need for Cognition \([\text{NFC}]\) and lower levels of framing-effects. However, these findings failed to be corroborated. Higher NFC is correlated with more consistency in choices (i.e. if they are presented with two versions of the same problem, higher-NFC people tend to make a choice in the second problem that is consistent with their choice in the first one); but higher NFC does not diminish susceptibility to framing effects (i.e. for high-NFC people their choice in the first problem is just as susceptible to framing effects as that of lower-NFC people). So insofar as NFC tracks greater use of reflection, it does not seem higher reflectivity lowers framing effects.

But there still is room for doubt. Need for Cognition is based on self-report, so it is more susceptible to inaccuracies than a direct measure of people’s cognitive dispositions. In this sense, the Cognitive Reflection Test \([\text{CRT}]\) seems to yield a more precise measure of dispositions towards reflection. There has in fact been some evidence suggesting that the CRT is correlated

\[39\] “Need for Cognition” \([\text{NFC}]\) is a measure of people’s disposition to engage in, and enjoy, reflective thinking about problems. The measure is based on self-reported agreement with statements like “I find satisfaction in deliberating hard and for long hours” and “Thinking is not my idea of fun” (Cacioppo & Petty 1982; Cacioppo et al. 1984).

\[40\] Smith & Levin (1996) initially suggested higher NFC scores correlated with lower susceptibility to framing effects, but this failed to be replicated by Levin et al. (2002) and LeBoeuf & Shafir (2003) among others. The latter study, however, found that participants with higher NFC tended to give more consistent answers across different framings of the same problem, even though their first answer was just as susceptible to framing effects as that of people with lower NFC scores. So higher NFC seems to be correlated with greater consistency, but not with lower framing-effect-susceptibility (which requires that answers display not only consistency but also ‘resistance’ to loss aversion).

\[41\] Give the following problem a try:

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? ___ cents.

This is a part of Frederick’s (2005) Cognitive Reflection Test, which measures how likely people are to override intuitive, but wrong, answers (in this case $.10), and reflectively generate the right answers ($.05). Most of us score low in the CRT. Unlike tests measuring optimal performance (like IQ tests or the GRE), the CRT measures typical performance (Toplak et al. 2011).
with framing-effect avoidance, but it is far from conclusive.\footnote{Frederick himself (2005, 32–34) showed that when making risky decisions low-CRT participants were risk-averse for gains and risk-seeking for equally large losses, whereas high-CRT participants were not. This by itself does not reveal anything about framing effects directly (the gain and loss problems in Frederick’s study were not identical); but it does show that the value function seems to differ among low-CRT and high-CRT groups.\textsuperscript{42}} On the other hand, there are some strong reasons to doubt high CRT scores are correlated with ability to avoid framing effects. In fact, studies of philosophers’ cognitive dispositions provide evidence against the correlation between CRT and framing-effect avoidance.

People trained in philosophy have been found to score higher on the CRT than those without philosophical training.\footnote{Cokely and Feltz (2009, 357) report a positive correlation between the number of philosophy courses taken and CRT scores among a group of more than 200 undergraduate students. Livengood et al. (2010) corroborate the philosophy-reflectivity correlation. In their study, people with no University-level philosophy training did worse at the CRT than people with some undergraduate philosophy training, and the latter did worse than people with some graduate-level philosophy training. The mean CRT score for the graduate-philosophy group tripled that of the no-philosophy group. Training in philosophy was predictive of higher CRT even controlling for educational levels.\textsuperscript{43}} Despite this (rather remarkable) fact, philosophers have also been found to be just as susceptible to framing effects as non-philosophers: to assess whether variations in framing-effects susceptibility correlated with variations in philosophical training, Schwitzgebel and Cushman (2015) presented philosophers, including professional philosophers, with the Asian disease problem. To increase reflective effort, they even encouraged a group to give “your reflective, considered responses”, and introduced a forced delay between question and answer. Moreover, their sample included philosophers who reported being experts in framing effects and loss aversion. To their surprise, they found that none of these groups (neither the}
group of those encouraged to reflect nor the group of self-declared experts) displayed lower susceptibility to framing effects than non-experts or non-philosophers.\footnote{Schwitzgebel and Cushman’s conclusion is in line with the position developed here: “Our results cast doubt on some commonsense approaches to bias reduction in scenario evaluation: training in logical reasoning, encouraging deliberative thought, exposure to information both about the specific biases in question and about the specific scenarios in which those biases manifest. Future efforts to minimize cognitive bias might more effectively focus on other means” (2015, 136).}

So people trained in philosophy tend to have unusually high scores in the CRT, but even professional philosophers (even those encouraged to reflect) seem to be just as susceptible to framing effects as non-specialist non-philosophers. All of which suggests that the amount of reflection, and the disposition to reflect, are not very promising debiasing tools when it comes to framing effects.

I do not mean to say that increases in reflection are absolutely ineffective in combatting framing effects. When it comes to numerical decision problems, numeracy is an especially powerful moderator of framing effects.\footnote{E.g. Stanovich & West (1998); Armstrong et al. (2006); Peters et al. (2006); Garcia-Retamero & Galesic (2010).} And, to be fair, some intensifications of reflection do seem to diminish framing effects: Sieck & Yates (1997) found that framing effects in the Asian disease problem were reduced when participants reflected on the problem for 50 minutes before making their choice, and were asked to explain their rationale. The point is rather that the situations in which a sheer increase in reflection diminishes framing effects are artificial and difficult to put into practice where they seem to be needed most (i.e. in situations where people with low numeracy levels face complex decisions under limited time constraints).

So evidence suggests that debiasing techniques of the ‘modify the decision-maker’ kind do not work when it comes to framing effects. In the next subsection I provide examples that show the ‘modify the environment’ approach can result more effective.

5.3.3. Better debiasing strategies for framing effects

Environmental debiasing techniques for framing effects

(A) Dispelling cognitive illusions through information design

When people have no other sources of credit available, they sometimes take payday loans. These work not according to an interest rate, but according to a fixed amount (e.g. 15 dollars for every 100 dollars lent). The borrower has to pay the loan back on her next payday—hence the name. By U.S. law, the payday lenders must inform their clients of the loan’s annual percentage rates,
which tend to be dramatically higher than other loans, like credit cards.\textsuperscript{46} But this information is usually buried within a long document, and many of the usual borrowers are not familiar with the APR concept anyway.

Bertrand and Morse (2011) tested whether a change in the way the information was presented could alter borrowers’ behaviour. Borrowers receive the cash and the loan documents in an envelope, so a group of borrowers were given the standard envelope, which highlights the date and amount they must pay (Fig. 5.5), and another group received an envelope showing how much the payday loan’s interest amounts to in comparison to a credit card’s interest after certain time periods (Fig. 5.6). The study found that those in the altered-envelope group were 11% less likely to borrow from the payday lender again in the four months following the intervention.

\textsuperscript{46} A payday loan’s annual percentage rate [APR] is usually around 400%. For comparison, the median microfinance loan has a 30% APR, and goes up to only 85% (Bertrand & Morse 2011, 1866).
11% may not seem much, but one must bear in mind that 76% of the payday loans are from old customers who take yet another loan within the two weeks immediately after they have paid the previous loan (Parrish & King 2009). The interest rate is so high that repaying the loan often requires a large portion of the next cycle of income, so that borrowers must take out another loan soon after, to pay for other necessary expenses. In this context, the fact that 11% more people were able to break the payday lending cycle is quite significant, especially considering the small change in the environment that the envelope’s design represents.

Bertrand and Morse claim the second envelope was an attempt to apply a number of behavioural principles to information design. The envelope re-frames the choice problem by making more accessible to the borrower the information about how the loan adds up over time, thus providing a broader temporal frame from which to make sense of the choice situation. It also provides a point of comparison to help the borrower in clearly assessing the magnitude of the payday loan’s interest rate. This design alteration sought to make it easier to overcome the illusion of present bias, i.e. our tendency to over-weight immediate gains and under-weight future losses (§5.2.4D). The simple envelope alteration managed to make the loan’s long-term costs more available to reflection, because it made them more intuitively salient.

A similar phenomenon occurs in the realm of medical decision-making. Informed consent laws require that patients be informed about risks before implementing a treatment. But the framing of risks can alter patients’ perception of such risks, especially when information is number-based and the patients have low numeracy levels. Garcia-Retamero and Galesic (2010) showed a great framing effect in low-numeracy participants when they were asked about the
level of risk of a procedure in negative terms (“Mr. Roe needs surgery: 9 in 1,000 people die from this surgery”) as opposed to positive terms (“Mr. Smythe needs surgery: 991 in 1,000 people survive this surgery”). In their study, however, when the problem information was accompanied with some visual aid (see Fig. 5.7), the framing effect almost disappeared.

Figure 5.7: Two visual aids found effective in debiasing framing effects (Garcia-Retamero & Galesic 2010)

Just like judgment errors in the Ebbinghaus illusion become more easily avoidable by adding an appropriate rectangle to the context, adding graphical representations to numerical decision situations also helps agents in constructing intuitive interpretations that bypass impending cognitive illusions (generated in this case by framing in terms of death or survival). As the researchers conclude, “[d]ifficulties with understanding numerical information often do not reside in the mind, but in the representation of the problem”.47

In a similar fashion, providing people with information about how their energy use compares to that of their neighbours tends to produce significant energy savings;48 informing people that a great part of their neighbours have already paid their taxes significantly increases tax compliance;49 simplifying official forms significantly diminishes errors in medical prescription deliveries;50 and timely text-message reminders to complete paperwork increases low-income students’ college enrolment,51 and timely messages. Anecdotally, Colombian grocery stores are legally required to display prices per unit (e.g. price per pound, gram or

47 Garcia-Retamero & Galesic (2010, 1323) (see also Fagerlin et al. (2005); Armstrong et al. (2006); Galesic et al. (2009); Garcia-Retamero & Dhami (2013)).
48 Allcott (2011); Behavioural Insights Team (2011, 18–26).
50 King et al. (2013).
51 Social and Behavioral Sciences Team (2015, 9–10).
millilitre) for every item; this simple information-design intervention makes the dozens of grocery-shopping decisions much easier and more precise.

These are cases of effective information design (i.e. information design that allow for points of comparison, simplify cognitive processing, and make relevant information available at the right time) that allow for the choice situation to be re-framed in ways that lead to increase in decision-making control, as revealed by increases in efficiency, accuracy, and inter-temporal consistency.

(B) Rebiasing toward better decisions

It is time to go back to the image with which this work began. Bateson et al. (2006) showed that consumers contributed almost three times as much money to a cafeteria’s honesty box when the price information was accompanied by images of watching eyes than when it was accompanied by images of flowers (see Introduction above). I take the Watching Eyes Effect to be a non-verbal framing effect, since the choice-problem was identical regardless of the accompanying image, but the image altered people’s preferences as to how much (or whether) to contribute.

The Bateson team has recently been taking it to the streets. One of their studies suggests that the Watching Eyes Effect is effective in deterring potential bicycle thieves from becoming actual bicycle thieves—or at least from doing so in locations with watching eyes (admittedly accompanied with some threatening words). According to the study, bicycle theft fell 62% in experimental locations, while it increased in control locations without eyes signs, which suggests they simply moved to where there were no eyes (Fig. 5.8).

Figure 5.8: Image of eyes looking after some bicycles (Nettle et al. 2012)

52 Nettle et al. (2012). The team has also found this effect regarding littering behaviour (Bateson 2013). However, Watching Eyes Effects seem to last for only a short time, and diminish as exposure increases (Sparks & Barclay (2013); see also the latter and Nettle et al. (2013) for meta-analyses).
This can be seen as a case of rebiasing: using cognitive biases (in this case the effect of being watched) to positively modify the agent’s decision-making process. The cases of making the most convenient option the default, and of framing medical decisions as losses or gains in order to boost preventive behaviours, also belong to this class. In a widely discussed example of this kind of debiasing strategy, Thaler and Sunstein (2008) propose a hypothetical school cafeteria in which healthier items are placed in more accessible and visible spaces, while junk foods and calorie-rich items are farther away from sight and reach. This is taken to practice quite literally in places like Toronto, where the cigarette rack is hidden from sight.

Thaler and Sunstein also discuss a very smart solution to a transit problem: Chicago’s Lake Shore Drive has S curves that make speeding very dangerous, since speeding cars can easily veer off the road. The city faced the issue by painting a series of horizontal strips that become closer together as they approach the dangerous curves. This gives the driver an illusion of going faster, which tend to make her slow down. In a similar vein, Philadelphia attempted installing 2D images of speed humps that gave the illusion of being 3D objects (see Fig. 5.9). This illusion would also get drivers to adjust their behaviour, even if only for the first few times.

Figure 5.9. Philadelphia’s virtual street hump
5.3.4. The power of environmental debiasing

The key to intelligent action is the automatic retrieval process.
—J. St. B. Evans (2008, 267)

Insofar as the above cases and arguments properly represent the field of framing-effect debiasing, it must be concluded that decision-maker debiasing strategies are rather unsuccessful when dealing with this particular bias, and environmental debiasing strategies are much more promising. What does this reveal about the nature of framing effects and the decision-making process?

Why is environmental debiasing more effective against framing effects?
There are several potential explanations for why increasing cognitive readiness (beyond basic levels of numeracy) does not seem to diminish framing effects. I will mention two likely ideas: first, thinking tends to seek confirmation rather than disconfirmation. Second, because of this, expertise or an intensification of reflection within a narrow frame leads to narrow thinking. To explain, consider the classical and widely studied Four-Card Selection Task:

Each of the boxes below represents a card lying on a table. Each one of the cards has a letter on one side and a number on the other side. Here is a rule: If a card has a vowel on its letter side, then it has an even number on its number side. As you can see, two of the cards are letter-side up, and two of the cards are number-side up. Your task is to decide which card or cards must be turned over in order to find out whether the rule is true or false. Indicate which cards must be turned over.\(^{53}\)

K  A  8  5

Most of us fail to find the right answer for this problem, and a large portion of respondents point to A, or A and 8. A is correct but insufficient (turning another card is necessary to find out whether the rule is true), whereas 8 is irrelevant (no matter what the 8 card has at the other side, it would not tell us whether the rule is true or false). Why do people tend to pick 8, if upon closer inspection it is irrelevant? Wason and Evans (1975) argue that this is because our card

---

choices are strongly influenced by a *matching bias*: we tend to select the cards that match what is explicitly mentioned in the rule (‘vowel’ and ‘even number’). After this bias influences our thinking, further reflection does little more than rationalize a distribution of salience already made at the intuitive level. In other words, we tend to reflect trying to *confirm* that the apparently relevant information is relevant, instead of trying to *disconfirm* it. (Similar phenomena seem to explain why [§5.3.2] the World Bank’s experts were especially susceptible to framing effects in cases in which the data did not match their prior theoretical commitments.)

To go back to the sequential dual-process model of decision-making sketched above (§5.2.3), it seems like we largely make up our minds in the first, intuitive phase of the process, in which we use intuitive processes to attribute degrees of relevance and irrelevance to different features of the situation, dissolving the situation’s felt tensions and completing the most apparent patterns. In the second phase, when reflection comes into the process, it tends mostly to look for information that corroborates what intuition has already highlighted as important. Escaping this tendency to confirm rather than disconfirm is very difficult, even for highly reflective thinkers. Once intuition has constituted a space of reflection, the fundamental features of that space are hardly questionable, even for reflection, because reflection tends to think ‘inside the box’ constructed by intuition. One may even say that the intuitive interpretation’s fundamental features—the box’s edges, so to speak—are mostly invisible to reflection: they belong to the realm of the un-reflected background, the situation’s boundaries that must remain un-thematized to function as such. Reflective processes need limited bits of information to function, and the automatic criteria employed by the intuitive selection processes (which include the heuristics and biases, and generate framing effects among others) are mostly hidden from reflection itself. So increasing reflection (either synchronically or diachronically) seems to be of very little use in revealing the edges of the box within which reflection takes place.

Framing effects are among the consequences of having a ‘reflection box’ that is too narrow for the present task, since some of the situation’s relevant aspects are left out, and too broad in another respect, since some irrelevant features are put in. Precisely because some relevant features are outside of the reflection box, they are not available to reflection; so it is difficult to conclude from inside the box that the box is too narrow. Often the cognitive failure in framing effects is not realizing that we are being subject to framing effects in the first place.

To escape that narrow thinking, it is not much help to think harder, because we will tend to think narrowly. What we need instead is to broaden our thinking box. And that is achieved by re-framing the situation, so that the set of features that become more intuitively salient captures more of the features that are relevant, and less of the ones that are not. This is
what happens in the cases of information design interventions (§5.3.3A) when people think about a payday loan’s costs in the longer term and in comparison with credit cards, or when they think about an operation’s riskiness with the aid of graphs. These are cases of environmental debiasing through information design: altering our environment so that the way information is presented leads to more useful intuitive interpretation, and thereby to more fruitful reflection. Cases of rebiasing (§5.3.3B), on the other hand, seem to work remarkably differently.

**Information design and re biasing are quite different strategies**

For another case of debiasing through information design, recall the Card Selection Task. It is quite hard to reach the correct solution once we are captured by the wrong answer’s allure. But it may be easier if the problem is re-framed. Consider this reformulation:

In its crackdown against drunk drivers, Ontario law enforcement officials are revoking liquor licenses left and right. You are a bouncer in a Toronto bar, and you’ll lose your job unless you enforce the following law: “If a person is drinking beer, then he must be over 19 years old.” The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of the card tells what a person is drinking and the other side of the card tells the person’s age. Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking the law. (Cosmides 1989, 192 [slightly modified])

The problem is logically identical to the original Card Selection Task, but as you can probably tell, the correct answer is much easier to grasp (drinking beer & 16 years old). Both versions of the problem are essentially identical, so the equivalent solution (A & 5) is correct for the original. In Cosmides’ study only 4–25% of the participants answered the original version correctly, whereas 75% gave the right answer to the ‘social norms’ version. This can arguably be seen as another case of environmental debiasing through information design.54

When it comes to debiasing framing effects, environment-modification techniques seem to work better than agent-modification techniques, because it seems to require thinking within

---

54 For discussion about why the social-norms version is more intuitive than the original, see Cosmides & Tooby 2013, 215–218.
broader frames, much more than thinking harder inside the narrow frame. The key is in modifying the first, intuitive phase of decision-making, rather than intensifying the second, reflective phase. This is ultimately good news from a practical standpoint, because changing an envelope’s design is much easier and less costly than training people in a cognitive skill or having them engage in intense reflection.

So far I have focused on the cases of environmental debiasing through information design; but these are the exception rather than the norm. Environmental debiasing is more widely known in its rebiasing variety. A crucial difference between an information design and a rebiasing intervention is that the former seeks to loosen the grip of bias by boosting the agent’s cognitive powers (by making the information’s format more amenable to those powers), whereas the latter seeks to use the agent’s biases in her favour. Given this difference, rebiasing is not sensu stricto a de-biasing tool; it may be more precisely portrayed as a bias restructuring technique. It does not seek to cancel biases, but rather to reorient them, so that they are aligned with the agent’s (or the collective) interests, idealized in some way.55

Another crucial difference between information design and rebiasing concerns their degree of accessibility. If participants in the Watching Eyes experiment were asked why they chose to give the specific amount of money they placed into the honesty box, they would surely not mention the images of eyes as important influencing factors. The same would apply for the examples in §5.3.3B: the rebiasing intervention would probably not be registered as significant in the decision-making process. By contrast, if the participants in the payday loan intervention who did not take another loan were asked why they decided not to do so, they would probably mention the information from the envelope as influencing their decision. Likewise, those who judge the Ebbinghaus circles to be equal in the altered version [Fig. 5.2] would probably mention the rectangle-based comparison as influencing their decision. So the effects of information design on the decision-making process seem much more accessible than those of rebiasing interventions.

One more crucial difference between information design and rebiasing concerns which part of the decision-making process they are intended to tackle. Information design tackles the initial phase: it seeks to diminish the grip of the cognitive illusion that generates the framing effect; rebiasing tackles the final phase: it seeks to improve the behavioural outcomes by steering the agent’s decision by biasing decision-making. A successful information-design intervention diminishes the grip of biases, thus increasing control over decision making; a successful rebiasing

---

intervention restructures biases, thus improving the action’s outcomes (by increasing some measure of gain or benefit) sometimes to the detriment of decision-making control.

Because rebiasing interventions restructure rather than counter biases, because they tend to be inaccessible to the agent, and because they are designed to improve behaviour rather than decision-making, there is legitimate concern that rebiasing may undermine agentive autonomy. On the other hand, information design strategies seem much less autonomy-endangering, because they are truly debiasing in nature, more thoroughly accessible by the affected agents, and designed to improve decision-making. If this is so, although both environmental debiasing tools seem effective against framing effects, information design interventions seem preferable than rebiasing since they are clearly less open to accusations concerning manipulation and autonomy (§5.2.4).

This leads to an answer to the section’s main question. Framing effects reveal the existence of a sort of decision myopia that poses risks to our control over decision-making processes and makes us susceptible to manipulation. What are the best strategies to counter framing effects? Evidence suggests that environmental debiasing is more successful at framing-effects debiasing than decision-maker debiasing; and reflection shows that among environmental debiasing techniques, information design is much better at enhancing autonomy and avoiding manipulation than rebiasing techniques.
§5.4. Control over decision-making and the theory of action

In this section I connect the study of framing effects to the broader questions about agentive control and the debate between intellectualism and anti-intellectualism that has been the topic of previous Chapters. I argue that (§5.4.1) the nature of framing effects suggests the process of intention production is guided by an automatic-associative capacity to make intuitive sense of practical situations. Such capacity, which I call background control, is a kind of automatic control that makes reflective, top-down control possible. This represents a return to anti-intellectualism, in the sense that reflective control turns out to depend on a prior, automatic kind of control, lacking which reflection is unable to produce intentional actions.

That said, (§5.4.2) framing effects also show that our background control capacities are severely limited—more so than anti-intellectualists usually admit—, and that we need the aid of propitious environments to aid intuition in avoiding cognitive illusions and making proper sense of practical situations and decision problems. Thus, (§5.4.3) in order to overcome decision myopia and enhance our autonomy, we must engage in reflectively designing our practical environments so as to empower background control.

5.4.1. Anti-intellectualism and the phenomenon of the background

Expanding the model of agency

By the end of Chapter 4, it seemed like intellectualism had the upper hand in the debate about action and agentive control. Reflective processes turn out to be necessary to explain all normal cases of intentional action, including expert, skilled, and even habitual sensorimotor action. The resulting picture is a hierarchical model in which action stems from reflective processes that constrain and inform automatic-associative processes by reference to an intention (Fig. 5.9).

Figure 5.9: An intellectualist model of human agency.
This model, however, leaves an obvious question open: where do intentions come from? How are they formed? This Chapter provides some elements toward an answer. If, as I have argued, we can learn about intentions (broadly construed) by studying decisions (broadly construed), and intention-formation processes by studying decision-making processes, then the hierarchical model of action in Fig. 5.10 can be complemented with a model of the decision-making process as it has been described in this Chapter (Fig. 5.10). The resulting model adds the intuitive sense-making processes necessary to produce the initial interpretation of any practical situation by attributing degrees of relevance and salience among its features. The upshot of this intuitive process (either mediated by reflective deliberation or not) is a decision, i.e. an intention.

![Figure 5.10: Adding decision-making to the human agency model.](image)

I propose now that a phenomenology-inspired notion of the background is useful in understanding the overall structure of the decision-making process, and the kind of control that we seem to have over it.

**Background control**
The dual-process model sketched above (§5.2.3) posits that the first phase of decision-making consists in an intuitive interpretation of the decision situation based on automatic-associative
processes. Said processes attribute relevance to some of the situation’s features—which makes them salient and available to reflection—and deem the others as irrelevant, making them largely inaccessible to further processing. But how do we intuitively assign relevance?

To answer this question, it is useful to introduce the concept of the background, which I understand as the broad sense-making toolkit composed of cognitive capacities, habits, bodily skills, intuitive heuristics, familiar tools, linguistic and communicative resources, and social relations, that an agent can automatically draw upon to make sense of a given situation. An agent’s background is structured partly by her body, her evolutionarily-structured cognitive structures, her accumulated experiences (stored as automatic associations), and her familiar practical environment and institutional settings. It is therefore potentially open-ended, and is a condition of possibility of any understanding of a particular situation. The background ‘qua background’ remains necessarily outside of reflection, but provides the agent with tools she needs to engage in reflective cognition, because it enables her to produce an initial automatic-associative interpretation of the practical situation she finds herself in.56

This initial interpretation is successful insofar as it presents as relevant what is really relevant in the situation for our current practical concerns. I call this capacity to correctly attribute relevance ‘background control’. It is an aspect of the automatic control whose reflection-independent normativity has been defended by anti-intellectualists (§2.1.2) and explored in Chapter 3. Background control belongs to the cognitive abilities in charge of the intuitive phase of decision-making.

How is the intuitive interpretation produced in a given practical situation? Upon facing the situation (or, in other words, a particular decision problem), some of the situation’s features appear to the agent, in her present state, as affordances (i.e. potentialities for action provided by the practical situation in connection with the agent’s background). But each situation has a great amount of potential affordances (e.g. a light switch affords turning the lights on; a water bottle affords grabbing and drinking; a book affords reading; a basketball affords dribbling), and most of the situation’s affordances are not relevant in any given moment. Some affordances, however, appear as solicitations, i.e. as affordances demanding a response from the agent. Solicitations immediately attract us, and prime us to responding in a certain way to the situation. Through

56 The concept of the background has permeated the phenomenological tradition since Husserl, is also present in the works of Wittgenstein and Searle among others, and continues to be a crucial component of some anti-intellectualist approaches to cognition and action. As may be expected, there are a number of complex theoretical discussions surrounding the notion, which deserve a careful treatment that go beyond this work’s limits. My immediate aim here is to construct a broad interpretation of the background to apply it to the question of agentive control that is this project’s central concern. (For a brief historical review see Dreyfus (2012). See Kiverstein (2012), Rietveld (2012), and the essays collected in Radman (2012) for recent in-depth critical discussion.)
solicitations, a particular way of acting is “summoned” by the practical situation.\footnote{The distinction between affordances in general and solicitations in particular is drawn by Rietveld (2012), following Dreyfus & Kelly (2007). See also Cappuccio & Wheeler (2012).} Anti-intellectualists describe the relevance-attribution process in terms that are familiar by now (§2.1.2): solicitations are experienced as tensions in the agent–environment coupling that request alleviation, or as environmental patterns that require a completion from the agent. (E.g. a decision problem whose alternatives are framed as losses solicits a risk-seeking decision to try and avoid the losses; in the bat-and-ball problem the $1.10 and $1 pattern solicits the $.10 completion; and as the alarm wakes me up to go running the gogginess solicits that I go back to sleep immediately.) Tensions to be alleviated and associative patterns to be completed are endowed with a high degree of relevance, loaded with affective and motivational associations, and provided with a salience that makes them potentially accessible to further cycles of cognition.

Drawing the intuitive distinction between affordances and solicitations is a process to which many of the background’s aspects contribute. Simple affordances become powerful solicitations by an interaction between the agent’s currently active bodily, cognitive, and motivational states, her stored associative networks (which include habits, memories, and skills), the way the practical environment is structured, and the intuitive structures of practical cognition. The latter are of particular relevance for us now. I have previously argued (§5.2.4) that loss aversion, status quo bias, present bias, and hyperbolic discounting are among those structures of practical cognition that form part of our backgrounds. These cognitive structures thus partially determine which affordances remain such, and which ones are promoted to solicitations, thus attaining higher degrees of relevance.

This intuitive relevance attribution in decision-making (what I have called ‘background control’) makes reflective aspects of practical control possible. In particular, it makes the second phase of decision-making possible by providing reflection with a finite set of environmental features to work with. Given our strict working memory limitations (§3.2.2), we would be unable to reflectively engage with the practical situation in its original richness of detail. Intuitive interpretation makes it possible for us to reflectively engage with a select subset of the situation’s features, thus making decision-making’s reflective phase possible (cf. Fig. 5.10) whenever the intuitive phase is not sufficient to produce a decision on its own. Thus, further downstream, intuitive sense-making makes intention-structured, top-down reflective control possible, by making intention production possible in the first place.

So we expand our agency model again, to include the background and solicitations (Fig. 5.11). The lines between the stages of action production are dotted to represent that they are...
porous processes in which the agent draws diverse tools from her background toolkit. Crucially, the background includes internal tools (like cognitive capacities or associative skills and habits), but it also includes external resources (like tools and social relations). This is why the circle delimiting the background is itself also dotted: the background is porous and open-ended.

![Diagram](image.png)

**Figure 5.11: Adding the background to the model of human agency.**

### 5.4.2. The limits of background control

**Background control and framing effects**

Recall the phenomenon of utilization behaviour, the neurological condition that makes agents respond immediately to any affordance, having lost sensitivity to the broader practical situation (§4.5.2). In his discussion of the phenomenon, Rietveld (2012) holds that utilization–behaviour patients “respond to irrelevant affordances, which results in inappropriate actions” and, moreover, that they seem unable to distinguish between relevant and irrelevant affordances; to them, all affordances are solicitations. He contrasts utilization behaviour patients with normal agents, who are quite good at intuitively drawing the affordance-solicitation distinction, and react appropriately when they have made a mistake in drawing the boundary between them (e.g. when they have made an awkward remark in conversation). By allowing for a contrast with the normal case, utilization behaviour
“makes clear that our normal everyday comportment is very well attuned to our personal interests and social context”. We, normal agents, are remarkably responsive to “real relevance”.

This is not entirely right. Framing effects, and the decision myopia that stems from them, prove otherwise. Not to overstretch the point, normal human agents are indeed quite good at making intuitive sense of the practical situations they find themselves in, especially if the situations are familiar to them, and they can use their background of internal and external tools to attribute relevance. That said, even in familiar contexts, normal agents are susceptible to misattributions of relevance, in large part because of the intuitive cognitive structures like present bias and loss aversion that make us susceptible to the cognitive illusions this Chapter has explored. These structures, which are parts of our background toolkit like any other, have the potential to make affordances into solicitations when we would be better off ignoring them, and to have us ignore affordances that should be solicitations. All of which may ultimately lead to severe errors of choice that threaten to undermine our autonomy. So we are more similar to utilization-behaviour patients than Rietveld and other anti-intellectualists may have warranted.

It follows that our background control capacities have severe limitations in certain aspects of the practical world (like risks and losses or satisfaction delays).

**Environmental debiasing is crucial in overcoming decision myopia**

The discussion of which debiasing strategies may help us avoid framing effects (§5.3) suggests that the limitations of background control cannot be corrected ‘from within’, by using the internal background toolkit; but rather that these limitations can be bypassed to the extent to which we add external tools to the background, which allow us to correct our intuitive blindspots. The solution to limits of background control can be found in environmental debiasing.

Sometimes when I wake up I do not feel like running. It is very easy to find reasons not to (“I didn’t get an optimal amount of sleep”, I say to my self, or “Tomorrow morning I’ll have more free time”). In these situations, it is quite helpful that I have made a commitment with my girlfriend that we would both run together. Thus, when I feel like not running she steers me toward sticking to the plan; and when she has trouble waking up I can annoy her until she is awake. This is environmental scaffolding, or environmental debiasing, in action. We would be much more myopic decision-makers if we could not resort to social relations and joint plans, timely reminders (in the form of calendars, to-do lists, or text messages), and the ability to
design our environments so as to make some options more immediately salient than others, some options easier to bring about than others.\textsuperscript{58}

This is an aspect in which Aristotle’s theory of action was close to being right, but not entirely right. He did notice, like the behavioural sciences have confirmed, that we are naturally more prone to immediate pleasures and less hard-working than we should (§1.2.2C), and in that he seems to have been right. But he proposed that the solution to this problem was acquiring the right character by means of a careful habituation that would make us able to endure and enjoy effortful activities (§1.3.1). In other words, he thought that the solution to our decision myopia was to be found in decision-maker debiasing. Now it seems to be clear that this is not the right way to tackle the problem: the limits of background control cannot be expanded significantly by training and skills. We rather need to offload the cognitive costs of decision-making to environmental scaffolds that allow us to correct our myopic attributions of relevance. Aristotle did notice that social practices and political institutions were required for individuals to achieve control over their goal-formation processes (§1.4.2); but he thought that this was because careful habituation was necessary to develop inner psychological capacities of rational care for one’s self. This Chapter’s argument suggests that Aristotle was right in considering social and political institutions crucial in attaining autonomy, but that not so much in thinking that this was due to the need to develop psychological capacities. We rather need social and political resources because autonomy must be environmentally scaffolded.

\textit{Scaffolded autonomy}

The coherence and the problem-solving power of much human activity, it seems, may be rooted in the simple yet often-ignored fact that we are the most prodigious creators and exploiters of external scaffolding on the planet. We build ‘designer environments’ in which human reason is able to far outstrip the computational ambit of the unaugmented biological brain. Advanced reason is thus above all the realm of the \textit{scaffolded} brain: the brain in its bodily context, interacting with a complex world of physical and social structures. These external structures both constrain and augment the problem-solving activities of the basic brain, whose role is largely to support a succession of iterated, local, pattern-completing responses.

—A. Clark: \textit{Being there}

\textsuperscript{58} This is a crucial tenet in Heath and Anderson’s (2010) argument for the extended will.
The study of framing effects has provided evidence to support the extended-mind thesis that cognitive processes are a dynamic coupling between agent and environment. Cases in §5.3.3 show that different environmental settings yield different cognitive decision-making processes. More strongly, they show that different environmental arrangements yield different degrees of control over decision-making processes. This strongly suggests that proper environmental arrangements (what Andy Clark calls cognitive scaffolding [see §2.1.2F]) extend our cognitive capacities, and make us able to tackle problems that we would otherwise be unable to solve. Examples of cognitive scaffolding are calendars, reminders, and to-do lists, diagrams and graphs, interpersonal commitments, and the underlying linguistic and numerical capacities that make these things possible.

An interesting aspect of this Chapter’s argument in relation to the extended-mind literature is the following. Usually reflective capacities are the ones that are in need of scaffolding. The canonical examples of pen and paper, or calendars and to-do lists, are meant to illustrate how environmental scaffolding can extend our working memory capacities. But the use of cognitive scaffolding in relation to framing effects reveals that our intuitive capacities are also in need of scaffolding. We need environmental debiasing tools in order for intuition to be able to surpass its relevance-assignment limitations. So not only advanced reason is the realm of the scaffolded brain, as Clark says. Unbiased intuition is also the realm of the scaffolded brain, and designer environments are necessary also to expand background control beyond its structural limits.

5.4.3. Conclusion: Indirect intellectualism and its implications

Indirect intellectualism

This is a summary of the argument so far: I have claimed that (§5.1.1) top-down reflective control is not sufficient to account for the phenomenon of basic agency, because this requires an account of the control we have over intention-formation (what we have called autonomy). This is mainly because we would not call someone an agent if she lacked all control over the intentions she executes (recall the case of the hypnotized patient, or the scopolamine victim). Relying on the psychology of decision-making, I have argued that (§5.4.1) we control our decision-making processes by means of an automatic kind of control, which I have called background control, i.e. the ability to use our background toolkits to distribute relevance among the features of a practical situation. But research on decision-making also reveals that (§5.2.4) background control has a series of limitations: we can fall into cognitive illusions, which generate framing effects and are able to lead us into decision myopia: the tendency to focus too much on avoiding (smaller or more
proximate) losses, thereby losing sight of (greater or more distant) gains. Decision myopia poses risks to autonomy, because it can lead us to distort our preferences, thus perverting the decision-making process, and makes us susceptible to manipulation.

What can we do to shield autonomy from decision myopia? Framing-effects debiasing studies show that (§5.3.2) we cannot use reflection to correct decision myopia, neither by intensifying it at the moment of deciding nor by developing more reflective competences and dispositions. However, these studies also show that (§5.3.3) we can use reflection to correct decision myopia by engaging in environmental debiasing.

The emerging picture of agentive control is that, on the one hand, reflective control is necessary for the production of intentional action; but, on the other, automatic background control is indispensable for the production of intentions. And whereas reflection cannot directly ensure that we are in control of intention-production, it can protect and extend our background control capacities by means of environmental debiasing. This leads to what I would like to call an indirect intellectualist view of action. It is intellectualist since it accepts the main tenet of intellectualism, namely, that reflective processes are necessary for the production of any full-fledged action. But it is indirect in the sense that reflection can contribute to the control of intention-formation only indirectly, by building the environmental scaffolding required for the proper functioning of automatic background control.59

If correct, indirect intellectualism has several important practical implications for our conception of what we should do, both individually and collectively, to maintain and increase control over our behaviour.

**Practical implications**

(A) An externalist approach to agentive control

The cognitive and behavioural sciences have left us a great philosophical task with its discovery of how widespread automaticity is in all aspects of everyday action, from decision-making to bodily action and social interactions. This work was inspired by the looming suspicion that we are no more than automata in most of what we do, and that our feeling of intentional control is not much more than an illusion. I am happy to report that the examination has provided positive results: we do have effective, albeit limited, control over our actions, and there are ways in which we can strengthen and expand said control.

59 For more on indirect intellectualism, and a comparison between this view and Aristotle’s indirect intellectualism, see the Conclusion below.
A crucial practical implication of this research is that we must look for solutions to problems of control outside rather than inside of the brain. If there is an illusion about agentive control, it is that it resides solely in the head, or in consciousness. Looking for it in there is part of the reason why it seems not to exist. A ‘mental muscle’ understanding of control suggests that we become more autonomous agents by thinking hard, that we develop a strong will by enduring temptations, thus building up mental stamina and becoming stronger over time. This may be true to some extent and with respect to some cognitive biases, but the present Chapter has offered evidence that when it comes to some of the deepest limitations of agentive control, we protect and increase our autonomy more effectively by building control into our practical environments, organizing the flow of action in ways that dissolve cognitive illusions or use our cognitive biases in our favour. There is no inner sanctum of individual choice, so our autonomy needs external tools and environmental design in order to extend the realm of decision-making in which it can function properly.

This externalist view of autonomy and agentive control has significant implications for how to build strategies of self- and community improvement.

(B) Environmental debiasing as a public policy issue

From the indirect intellectualist perspective, agentive control is not a property that we human beings have in virtue of having some internal cognitive capacities. Such capacities are necessary conditions for successful agentive control over intention and action production, but they are not sufficient, given the limitations of both our intuitive structures of practical cognition, and our reflective capacities of executive control. Given this, agentive control is to some extent an achievement, and achieving it depends on building and maintaining the environmental structures that allow for us to extend our cognitive capacities beyond what is made possible by our bodily cognitive systems.

Some of these environmental structures are a matter of private responsibility: we are in charge of structuring our living spaces and caring for our relationships, for instance. But some of these environmental structures are public resources: language is a crucial example, but there are also social norms, which often establish the status quo of interaction; public institutions, which set default options and design crucial information about services and obligations; and technological instruments, through which our capacities of access to information and its manipulation are extended. These are examples of shared, public resources on which our capacity

---

60 I draw this idea of control externalism from Heath and Anderson’s (2010) externalist theory of the will, which in turn is a development of Andy Clark’s mind externalism.
for agentive control partially rests. Such resources are constitutive parts of each human agent’s background, thus empowering or diminishing her background control, which is a condition for her individual autonomy.

So individual autonomy rests partly upon shared, public cognitive resources. These resources (such as public services and technological tools) are unequally available to members of our societies, and this implies that the levels of individual autonomy that each one of us achieves partially depends on the distribution of such public resources. Thus environmental scaffolding becomes a matter of distributive justice. If this is so, then access to environmental scaffolding (particularly to the scaffolding that generates environmental debiasing) turns out to be a public policy issue: governments that are concerned with contributing to their citizens’ autonomous decision-making capacities have, to some extent, a duty to make environmental debiasing tools widely available.

The social and political implications of indirect intellectualism are worth detailed examination, a task that must be left for a future project. I close this Chapter by briefly returning to one of the debates that merit later careful discussion: the issue of debiasing and manipulation.

**Coda: Environmental debiasing and the problem of manipulation**

The application of behavioural and cognitive insights to public policy has become a trend, largely inspired by Cass Sunstein and Richard Thaler’s defence of the stance called 'libertarian paternalism'. This approach has been criticized for seeming manipulative in a way that threatens to undermine autonomy. Is libertarian paternalism a useful framework to justify the use of environmental debiasing techniques? Can the discussion above shed some light on the issue of manipulation?

**C** Debiasing and libertarian paternalism

Libertarian paternalism holds that choice architects should steer people toward better decisions by use of *nudges*, i.e. by manipulating aspects of the choice architecture that can predictably improve agents’ decisions and behaviour, on the condition that such manipulation must keep all options available and must not alter the incentive structure.\(^6\)

The approach is called *paternalism* because it endorses the governmental manipulation of choice architecture towards ‘improving’ the decisions of individual citizens. Yet it is also called *libertarian* because it is designed to preserve freedom: all options remain available and no option is rendered more expensive. Libertarian paternalism is attractive, among other things, because it

---

\(^6\) Thaler & Sunstein (2008).
suggests that interventions in the choice environment which are often cheap and hopefully freedom-preserving can steer people toward making better choices.

Any of the various debiasing strategies summarized above count as nudges. This presents a conceptual problem, since the strategies differ in many respects, as seen above. The problem is reflected in the broad range of things the authors are willing to call ‘nudges’:

Nudges are interventions that steer people in particular directions but that also allow them to go their own way. A reminder is a nudge; so is a warning. A GPS nudges; a default rule nudges. To qualify as a nudge, an intervention must not impose significant material incentives. A subsidy is not a nudge; a tax is not a nudge; a fine or a jail sentence is not a nudge. To count as such, a nudge must fully preserve freedom of choice. If an intervention imposes significant material costs on choosers, it might of course be justified, but it is not a nudge. Some nudges work because they inform people; other nudges work because they make certain choice easier; still other nudges work because of the power of inertia and procrastination. (Sunstein Forthcoming §11)

The criticism has been raised that it is not actually clear how libertarian paternalism is related to nudges: if anything that steers people toward a certain decision without altering incentives or prohibiting options counts as a nudge, then a GPS and rational persuasion count as such, but we would not call either a libertarian-paternalistic intervention.62 That said, all libertarian-paternalistic interventions do seem to require nudges, and indeed a number of governments around the world have begun to implement such interventions by designing nudges, like altering default options to increase the percentage of people enrolled in pension savings or organ donor plans; changing the way official forms are structured to increase compliance and diminish errors; carefully designing the wording of communications to increase the percentage of tax compliance; and many others.63

How do nudges relate to debiasing techniques? The word ‘nudge’ covers countless things from establishing defaults to using a GPS, so it seems safe to say that anything that counts as a debiasing intervention—either information design or rebiasing—also counts as a nudge. Given this, the debiasing taxonomy developed above may contribute to one of the ethical debates surrounding libertarian paternalism, namely to what extent nudges are worryingly manipulative.

---

62 See Wilkinson (2013) for this and other conceptual qualms with Thaler and Sunstein’s approach.
63 For reviews of recent libertarian-parternalist applied research and programmes, see e.g. Behavioural Insights Team (2014); Lunn (2014); World Bank (2015); Social and Behavioral Sciences Team (2015).
(D) Is environmental debiasing manipulative?

I have argued that decision myopia is a matter of public policy concern. Now, is libertarian paternalism the right public policy approach? There are many criticisms to it, both epistemological and ethical. Epistemological criticisms gravitate around the worry that choice architects have no reliable way of knowing what individuals’ interests and preferences are any more than individuals themselves do, and that thus libertarian paternalism often ends up replacing individuals’ own preferences and conceptions of wellbeing with those of choice architects themselves. Libertarian paternalism thus seems to rest on the unjustifiable assumption that choice architects can ‘fix’ decision-making processes because they have a privileged access to people’s own preferences—which they do not.64

One of the most pressing ethical concerns suggests that this approach is not libertarian enough to avoid the problems of classical paternalism. Specifically, libertarian paternalism’s manipulation of choice architecture may be too manipulative to sufficiently preserve individual autonomy.65 As seen above (§5.2.4), the concern about manipulation goes beyond the manipulation’s beneficial outcomes. Likewise, paternalistic manipulation would be objectionable even if public officials were somehow able to gain accurate epistemic access to wellbeing or people’s ‘true preferences’. In Darwall’s words, the underlying problem with paternalistic manipulation is “primarily a failure of respect, a failure to recognize the authority that persons have to demand, within certain limits, that they be allowed to make their own choices for themselves”.66 The problem is, in other words, that of undermining autonomy by manipulating decision-making.

However, this criticism does not apply equally to all nudges. Information design nudges, as seen above, enhance rather than endanger autonomy, and seem to help in avoiding manipulation, rather than manipulating themselves; the same does not apply to rebiasing nudges (§5.3.4).67 That said, manipulation and autonomy are not contradictory terms: some kinds of manipulation are compatible with autonomy, and can even enhance it, in the right circumstances. Many would consent to campaigns that print terrifying images in cigarette boxes, or that keep them hidden, even though these constitute manipulations, on the grounds that these rebiasing effects make self-controlled behaviour easier; and if consent is given, no harm to

64 For criticisms along these lines, see White (2008); Rebonato (2014); Waldron (2014).
65 See Anderson (2010); Wilkinson (2013); and the works by White and Rebonato in the previous note.
67 Cf. Sunstein (Forthcoming, §87).
autonomy seems to ensue. Moreover, since time and cognitive resources are limited, we may often prefer to rely on well-designed rebiasing nudges that determine appropriate defaults and place useful cognitive illusions in our ways, so that we can save the effort of choosing; in other words, we may ‘choose not to choose’. 

These considerations, however, are only the tip of the iceberg. When would it be safe for governments to remove the burdens of choice? What would be the difference between a nudge and a non-paternalistic policy if consent is given? And how could consent be given on an issue surrounded by cognitive illusions? Would it be better to encourage active choice, despite preferences to bypass choosing, on the grounds that through practice people can become better at decision-making? These are among the questions that a more thorough investigation of the ethics of behavioural policy making must tackle.

---

69 Sunstein (2014).
Conclusion

This work opened with a couple of central questions: What kind of agency and control do we humans really have over our behaviour, if most of it is automatic? How does it work, and what are its limits? The Chapters have developed a theory of control that synthesizes several lines of empirical research, has important practical implications, and advances theoretical discussions about agency in several directions. In what follows I unpack some of the theory’s virtues and achievements.

Control: Its kinds and its limitations

Human agentive control is split into two main kinds: reflective control, which is grounded on the use of executive functions (i.e. the capacities to coordinate attention and memory for the attainment of complex goals); and automatic control, which consists in the error-detection and correction mechanisms of the mind’s association-based intuitive systems. Automatic control can in turn be subdivided into coping control (i.e. the ability to produce particular cognitive or sensorimotor routines) and background control (i.e. the capacity to make intuitive sense of a practical situation).

Each kind of control has its limitations: the limits of reflective control are equivalent to those of the executive functions. Coping control is limited by the fact that automatic-associative processes are syntax-insensitive, and can sometimes conflict with higher-order reflective intentions and goals, and can be inaccessible to consciousness. Background control, on the other hand, is susceptible to framing effects and other cognitive biases, which may lead us to interpret irrelevant environmental features as relevant, and vice-versa, leading to myopic decision-making processes. However, all these kinds of control, although limited, are capable of being extended by background tools and environmental scaffolding that help reflection extend its calculative and memory capacities, and also aid intuitive control in avoiding or neutralizing the cognitive illusions that lead to biases and loss of autonomous decision-making.

So what kind of agency do these types of control provide? Concerning the debate around intentional action, I argued (against anti-intellectualism) that all intentional actions, even of the skilled and habitual kind, require the reflective, top-down modulation of attention in relation to a higher-order goal; without this modulation, behaviour can too easily lose the minimal amount of coherence and unity. This entails that there are no actions that are purely automatically controlled. But I also argued that the view of agency must be expanded to include an aspect that intellectualism usually avoids: the production of intentions and goals. Looking at the
phenomenon of decision-making as a model for intention-generation in general, I showed how the production of goals depends on basic intuitive processes that make sense of the practical situation by attributing degrees of relevance to each feature of the situation. Since only the relevant features become available to reflection, and they are available under the description intuitively given to them, the decision-making process, even if it involves reflection, depends on the prior automatic processes that interpret each situation.

The ultimate upshot from this is the view that individual agentive control is largely dependent on social and environmental support structures, that allow for reflective and intuitive extension of control (Chapter 5). Accepting this implies returning to an Aristotelian position on human agency, in the sense that according to Aristotle individual agency rests on a context of public care, without which the former would be drastically reduced.

*Indirect intellectualism, old and new*

Indirect intellectualism is a way of assimilating dual-process theory of cognition and action. Having cognition split into two process types implies that there are some functions that the reflective type of processing cannot accomplish itself. Particularly, in the account I have developed, reflection is unable to produce the originary interpretation of the practical situation that sets the stage for decision-making and action. Given this, reflective control’s participation in guiding and enhancing those first steps of decision-making is mainly indirect, by providing the means whereby the automatic sense-making processes can work smoothly.

There is a very close analogy between this indirect intellectualism and Aristotle’s. For arguably his psychology was also a dual-process one: Aristotle conceives of a character-reason distinction that significantly parallels that between intuition and reflection. And his indirect intellectualism also originates in the cognitive division of labour: because an agent’s ends can be established only by character, reason’s role in this respect must be indirect: it contributes to establishing the appropriate goals by carefully guiding the habituation processes that produce a good character.

As pointed out earlier, a crucial difference between the two versions of indirect intellectualism is that Aristotle’s is much more internalist in comparison to agentive control than the new version I have defended here. Aristotle has a conception of human character as a feature of an individual that is much more stable than the decision-making capacities that current behavioural and cognitive sciences depict. Whereas Aristotle’s virtuous man is constantly virtuous throughout many different circumstances, a good decision-maker in many contexts may
nevertheless find choice situations in which cognitive illusions easily distort her decision-making processes. This explains why Aristotle places great emphasis on habituation, whereas I have argued we should pay serious attention to environmental design.

I have shown that, as an interpretation of Aristotle’s action theory, indirect Intellectualism preserves the benefits of both Intellectualist and anti-Intellectualist accounts while avoiding the pitfalls of each extreme: it preserves the anti-Intellectualist claim that goals are set by the non-rational part of the soul, while managing to explain how reason can still command and direct the non-rational part. This is the work’s main contribution to Aristotle interpretation.

Similarly, indirect intellectualism in the contemporary context preserves the crucial insights of both intellectualism and anti-intellectualism about intentional action. While acknowledging the crucial role of executive control of attention in action production, indirect intellectualism also incorporates the anti-intellectualist claim that much agentive control, and the primary sense-making relationship between agent and world, is pre-reflective, automatic, and intuitive. One of the virtues of indirect intellectualism for the current discussion is its ability to combine both views within a cohesive framework. Another one is its explanation of how reflective control can enhance automatic control of processes that nevertheless occur pre-reflectively: careful, reflective environmental design allows for automatic control to avoid the pitfalls of intuition generated by cognitive biases.

Finally, indirect intellectualism is sufficiently robust to dispel the paradox of automaticity introduced in Chapter 3 (§3.1). The pervasiveness of automaticity in everyday behaviour poses a threat to agentive control for two reasons: first, directly, because the structural biases in our intuitive systems of practical cognition can lead to decision myopia; and second, indirectly, because the capacity limits of executive functions can lead to distraction, slips, and other disunity problems in which behaviour is automatically controlled but lacks intention-based structure. But automaticity is also necessary for action, also in two ways: first because habitual and skilled action requires automatic coping control to match relevant inputs and relevant outputs; and second, because all action-production depends on intention-production, which depends on automatic background control. According to indirect intellectualism, the threats of automaticity can be solved by developing the right kinds of environmental scaffolding, which extend both our reflective and intuitive cognitive capacities. This extension is an ongoing task, whose continuation pertains to all of us, both individually and collectively.
## Bibliography

### Ancient works

**Aristotle**

- **APo** *Posterior Analytics*  
- **DA** *On the Soul*  
- **EE** *Eudemian Ethics*  
- **EN** *Nicomachean Ethics*  
- **GA** *On the Generation of Animals*  
- **HA** *History of Animals*  
- **MA** *On the Motion of Animals*  
- **Men.** *On Memory and Recollection*  
- **PA** *On the Parts of Animals*  
- **Phys.** *Physics*  
- **Pol.** *Politics*  
- **Rhet.** *Rhetoric*  
- **De sensu** *Sense and Sensibilia*  

**Diogenes Laertius**

- **VP** *Lives of Philosophers*  

**Homer**

- **Od.** *Odyssey*  
Plato

*Phaedr.*


Sextus Empiricus

*PH Outlines of Pyrrhonism*


Modern works


Behavioural Insights Team (2011). *Behaviour change and energy use*. Cabinet Office UK.

— (2012). *Applying behavioural insights to reduce fraud, error and debt*. Cabinet Office UK.


Bibliography — 228


Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). “All frames are not created equal: A typology and critical analysis of framing effects.” *Organizational behavior and human decision processes* 76:2, 149–188.


— (Forthcoming). *The myth of ‘Just do it’: Thought and effort in expert action*. Oxford University Press.


