Technology and the Human Minds

Keith Frankish

Abstract *According to dual-process theory, human cognition is supported by two distinct types of processing, one fast, automatic, and unconscious, the other slower, controlled, and conscious. These processes are sometimes said to constitute two minds—an intuitive old mind, which is evolutionarily ancient and composed of specialized subsystems, and a reflective new mind, which is distinctively human and the source of general intelligence. This theory has far-reaching consequences, and it means that research on enhancing and replicating human intelligence will need to take different paths, depending on whether it is the old mind or the new mind that is the target. This chapter examines these issues in depth. It argues first for a reinterpretation of dual-process theory, which pictures the new mind as a virtual system, formed by culturally transmitted habits of autostimulation. It then explores the implications of this reinterpreted dual-process theory for the projects of cognitive enhancement and artificial intelligence, including the creation of artificial general intelligence. The chapter concludes with a brief assessment of the risks of those projects as they appear in this new light.*

1 Introduction

Over the last 40 years, many psychologists have come to adopt some form of *dual-process* theory. Such theories hold that human cognition is supported by two distinct types of processing—a fast, automatic, unconscious type, and a slower, controlled, conscious one—which can yield different and sometimes conflicting results. The distinction corresponds to the everyday one between intuition and reflection, the former delivering spontaneous responses that just feel right, the latter more considered responses for which one can give some explicit justification. These types of processing are sometimes said to be associated with two brain systems, System 1 and System 2, the first evolutionarily ancient and largely shared with other animals, the latter more recent and distinctively human. Some dual-process theorists speak of our having two *minds*, an intuitive *old mind* (System 1) and a reflective *new mind* (System 2) (Evans 2010).

Such views have obvious implications for cognitive enhancement and artificial intelligence (AI). If we do have something like two minds, then the projects of enhancing and replicating human intelligence will each also assume a dual aspect. Dual-process views face some problems, however, and it is hard to see how System 2 could be modelled artificially. These problems, I believe, dictate a reinterpretation of dual-process theory, which pictures the two minds as levels of organization rather than distinct systems. The new mind should be seen, not as a brain system, but as a virtual one, formed by culturally transmitted habits which restructure the activities of the old mind. This reinterpretation helps to resolve some of the problems for dual-process theory and makes the project of artificially creating a System 2 mind somewhat more tractable.

In this chapter I shall explore these issues, explaining the re-interpretation of dual-process theory, looking at its implications for projects of cognitive enhancement and AI, and assessing the risks of those projects as they appear in this new light. I begin, however, by introducing the dual-process approach.

2 Dual processes

Dual-process and dual-system theories grew out of experimental work in cognitive and social psychology from the 1970s onwards (for on overview, see Frankish 2010). The theories were formulated in a series of important papers and books published in the late 1980s and 1990s (e.g., Chaiken and Trope 1999; Chen and Chaiken 1999; Epstein 1994; Evans 1989; Evans and Over 1996; Petty and Cacioppo 1986; Sloman 1996; Stanovich 1999; Stanovich and West 2000) and brought to a wider audience in several books published over the next decade or so (Evans 2010; Kahneman 2011; Stanovich 2004).

Many variants of the dual-process approach have been developed, differing in detail but agreeing on the fundamentals. A composite account incorporating the most common claims runs as follows. There are two types of processing ('thinking') involved in human reasoning, decision making, and social cognition: Type 1 and Type 2. Type 1 processing is typically fast, automatic, effortless, non-conscious, associative, parallel, high-capacity, and undemanding of working memory. It is highly contextualized, draws on implicit knowledge acquired from past experience, and delivers responses that may be adaptive in real-world settings but often deviate from rational norms, manifesting cognitive biases, stereotype effects, and emotional influences. Type 2 processing, by contrast, is typically slow, controlled, effortful, conscious, rule-governed, serial, low capacity, and demanding of working memory. It is more abstract, draws on explicit knowledge and learned rules of inference, and is more likely to deliver responses in line with normative principles. Type 2 processing is also linked to hypothetical thinking—evaluating candidate actions in imagination and simulating alternative perspectives and scenarios. For this, we must entertain 'secondary' representations, which are decoupled from the world and do not directly affect behaviour, and this is held to require Type 2 processing. Finally, the propensity to use Type 2 processing shows high individual variability and is correlated with measures of general intelligence.

Dual-process theorists differ as to how the two processes are related. Some see them as operating independently and competing for control of behaviour. Others adopt a *default-interventionist* model, according to which Type 1 processes supply rapid default responses, which can be intervened upon and overridden by Type 2 processes. On this view, Type 1 processes are also responsible for triggering Type 2 processing and for selecting information for it to use (Evans 2006; Kahneman 2011).

Dual-system theories propose a broader architectural basis for the two types of processing, which assigns them to different mental systems, System 1 and System 2. System 1 is taken to be composed of multiple subsystems, many evolutionarily ancient, all of which operate in a Type 1 way (e.g., Stanovich 2004). These include perceptual, motivational, and emotional systems, learning and conceptual systems (perhaps specialized for particular tasks, such as navigation, foraging, social cognition, theory of mind, and language), and procedures for learned skills practised to automaticity, such as reading and driving.[1] System 2, on the other hand, is thought of as a single, low-capacity system which can manipulate explicit representations in working memory. It is flexible, responsive to instructions, and uniquely human.

There is a mass of evidence for the dual-process picture, from three independent sources (for a summary and illustrative references, see Evans and Stanovich 2013). First, there is evidence from response patterns in reasoning and decision-making tasks. Typically, participants give one of two answers, the first intuitively plausible but normatively incorrect, the second less obvious but correct, and experimental manipulations can influence which it is. For example, time pressure leads to increased production of the intuitive answer (rather than random responding), whereas clear task instructions promote the normatively correct one. This strongly suggests that two different mechanisms are in play, one fast and intuitive, the other slow and reflective, each of which delivers a specific answer. A second source of evidence comes from work on individual differences. There is a positive correlation between tendency to give the normative responses on reasoning tasks and general intelligence, which is explained on the hypothesis that those of higher general intelligence have a

[1] The basic dual-system framework is compatible with a spectrum of views as to the nature of the evolved components of System 1, from ones which posit multiple domain-specific modules (e.g., Carruthers 2006) to ones which hold that learning is domain-general and that specialized systems are *cognitive gadgets* installed by cultural processes during individual development (Heyes 2018).

greater capacity to engage in and sustain Type 2 processing and to override intuitive Type 1 responses. (This is not surprising, given that Type 2 processing requires working memory, and working memory capacity itself correlates with general intelligence.) Third, there is neuroscientific evidence. Imaging studies indicate that different neural structures are involved in the production of responses associated with each type of processing, Type 2 responses typically following activation of prefrontal and frontal cortical regions that are not involved in Type 1 responding.

This dual-system view has a common-sense appeal, and something like it has been tacitly acknowledged for centuries (Frankish and Evans 2009). Many early modern philosophers agreed with Descartes in identifying the mind with the conscious mind, understood as an immaterial substance that is the arena of pure thought. But they also recognized that much human and animal behaviour occurs without conscious thought and must be supported by complex nonconscious mechanisms of some kind. (Descartes himself fully recognized this; Descartes 1984, p. 161.) The development of scientific psychology in the nineteenth century saw the gradual acceptance that these processes were genuinely mental, involving non-conscious perceptions and thoughts, operating independently of the conscious mind. More recently, with the development of the computational theory of mind and modern cognitive science, non-conscious processes increasingly took centre-stage in the explanation of human behaviour, with the conscious mind sometimes being demoted to the role of a rationalizer (Wegner 2002).

In the history of AI too, we can see implicit acknowledgement of the two-systems distinction. Early AI researchers focused on abstract reasoning and decision making, which they sought to model in computational terms, with the aim of creating artificial general intelligence. Lack of success in this project led many researchers to turn to a bottom-up approach, seeking to create embodied, robotic systems with specific behavioural competences (e.g., Brooks 1991; Steels and Brooks 1995). From a dual-process perspective, this was simply a switch of focus from System 2 to System 1.

3 Problems

Despite its attractions, dual-process theory has its critics (e.g., Gigerenzer 2010; Keren and Schul 2009; Kruglanski and Gigerenzer 2011; Melnikoff and Bargh 2018; Osman 2004, 2018). A common objection is that it is highly unlikely that the various features ascribed to each process (fast vs slow, automatic vs controlled, non-conscious vs conscious, etc.) align so neatly, excluding crossover processes that are, for example, fast but controlled. Critics also object to the suggestion that intuitive processing is always biased and reflective processing always normatively rational.

Dual-process theorists respond by clarifying the scope of their claims (e.g., Evans and Stanovich 2013; Pennycook et al. 2018). They explain that the features ascribed to each process are not all defining ones and that the core distinction can be drawn more simply. Evans and Stanovich identify autonomy (lack of attentional control) as the defining characteristic of Type 1 processing, and the use of working memory and support for decoupled representations as those of System 2 (Evans and Stanovich 2013; Stanovich and Toplak 2012). The other features commonly ascribed to each system are held to be merely typical correlates of these defining features. For example, Type 2 processing is typically slow and serial because it loads on working memory, which is a limited resource. As Evans and Stanovich stress, this allows for considerable variation in the mode of Type 2 thinking, since individuals may use many different procedures and strategies for manipulating explicit representations in working memory, reflecting their individual 'thinking dispositions' or 'mindware' (Stanovich 2009a, 2011). For the same reason, it is wrong to think that Type 2 processes always deliver normatively correct responses and that all cognitive errors are due to Type 1 processes. It is true that this pattern is often observed in experimental settings designed to create conflict between the two kinds of processing, but there is no reason to think that it is a universal one. Type 2 processing may often deliver incorrect or biased responses, owing to inattention, misunderstanding, or poor strategy (buggy mindware) (Evans 2006, 2007; Stanovich 2009b). Conversely, intuitive Type 1 processing may often deliver optimal responses, at least in favourable conditions.

Recently, theorists broadly sympathetic to the dual-process approach have raised more specific worries, especially about the relation between the two systems. These have prompted proposals for the revision or refinement of the framework, though without undermining the case for a qualitative distinction along the general lines proposed (De Neys 2018).

There is, however, another, more general problem I want to raise for the dual-process theory. It concerns Type 2 processing. What exactly is the *mechanism* by which this processing operates? Calling it a *system* implies that the reflective mind is a self-contained device, which takes inputs from System 1 but processes them using its own proprietary mechanisms. Theorists identify various components of this device, including working memory, explicit decoupled representations, and executive control processes, but these do not in themselves amount to a reasoning system. What is the engine that manipulates the explicit representations in working memory, in accordance with rules of inference or other procedures? Dual-system theorists are strangely silent on this.

There is a related evolutionary worry. If System 2 is a self-contained device, it must be an extraordinarily powerful one. We can turn our conscious minds to any problem. We can think about things distant in time and space and about abstractions and hypothetical scenarios. We can construct rational arguments, form and evaluate novel ideas, devise complicated plans of action, and much more. How and why did such a system evolve? Although the human brain is much larger than the brains of other animals, its evolution seems to have involved the addition of new specialist subsystems, such as ones for language, mindreading, and social cognition, and the enhancement of existing ones, rather than the installation of a completely new general-purpose reasoning system (Carruthers 2006). Indeed, it is hard to see what evolutionary pressures there could have been for the development of such a system. Having a capacity for flexible, abstract, rule-governed deliberation is advantageous in the modern world (a world that is largely the creation of our human minds), but it is hard to see why it would have been required in the ancestral environment in which our species evolved. Cognitive flexibility is certainly useful, but to build in general intelligence seems like a massively overengineered solution to any specific environmental challenges our ancestors might have faced.

4 Type 2 thinking as an activity

I want to suggest a reinterpretation of dual-process theory, which helps to address these problems.[2] The key idea is that some thinking is an intentional activity, something we *do*. The distinctive thing about intentional actions, as opposed to other bodily movements and processes, is that they are under voluntary control, responding to our beliefs and desires. We perform them because we *want* to—either because we enjoy them or because we believe they will further some goal we have. These reasons need not be consciously entertained. Most of our behaviour is unreflective: we walk, talk, drive, and go about our daily lives without giving much conscious thought to the reasons for our actions. But the actions are still intentional ones, directed to our goals and guided by our beliefs. The defining feature of Type 2 thinking, I propose, is that it involves performing intentional actions in this sense. Type 1 processing, by contrast, is a wholly automatic process, which occurs without our needing to do anything.

[2] This interpretation draws on suggestions by Dennett and Carruthers, among others (Carruthers 2006, 2009; Dennett 1991). For further explorations of the view, see Frankish (1998, 2004, 2009, 2018). It is possible that some dual-process theorists always intended this interpretation and that it is implicit in the characterization of Type 2 processing as *controlled*. If so, then the present proposal is more an explicitation than a reinterpretation.

How could reasoning be an intentional activity? Consider solving a long division problem with pencil and paper, following the procedure you were taught at school. We write out the numbers in a certain format, then do a series of simpler calculations, each step building on the previous one, until we arrive at the solution. This involves a series of actions—writing down various numerals in certain locations—which are performed with the goal of solving the problem. But how do we know which actions to perform at each step—which numerals to write and where? What is the reasoning mechanism that takes us from step to step, from one set of symbols to the next? The answer of course is that it is System 1. The answer to each subproblem comes to us intuitively, courtesy of automatic Type 1 processes. When we need to subtract two from seven, say, we just see that the answer is five, and write it down. Each step in the controlled, conscious procedure is driven by intuitive Type 1 processes which are neither controlled nor conscious. Indeed, the role of the procedure is precisely to break down a complex problem that we cannot solve intuitively into smaller problems that we can. The process is, we might say, one of deliberative *mastication*.

A similar process can be used to reason in a more exploratory way. Skilled mathematicians can combine various pen-and-paper procedures, supported by a rich intuitive understanding of the subject, to explore novel theoretical possibilities. Again, the manipulation of written symbols allows them to break down a complex problem into intuitively manageable chunks.[3]

How exactly does intentional reasoning like this work? It is useful to think of it as operating by means of what Daniel Dennett calls *autostimulation* (Dennett 1991, Ch. 7). In creating and manipulating external symbols we are cognitively stimulating ourselves, providing new inputs to our Type 1 mental processes. Our perceptual systems detect and interpret the symbols we create, and conceptual, emotional, and motivational systems get to work on the problem of how to respond (what to write next and where). These systems compete for control of motor systems, leading to a further action, which forms the next step in the sequence. We also create drawings and diagrams to help us solve problems and evaluate options. Think of making sketches to experiment with designs for a garden or for the layout of furniture in a room. Again, the process is autostimulatory. We sketch a design, examine it, and our autonomous mental processes generate an evaluative response. Perhaps the design looks ugly or unbalanced or just wrong somehow.

[3] The physicist Richard Feynman insisted that his notes were not a record of work done in his head but the very working itself. "No, it's not a record, not really. It's working. You have to work on paper, and this is the paper. Okay?" (quoted in Gleick 1992 p. 409).

But the most powerful means of autostimulation is speech. By talking to ourselves we can work our way through a tricky problem. We question ourselves ('Where did I leave the remote?'), guide ourselves ('That's the earth pin, so this must be the live'), prompt ourselves ('It begins with a T'), encourage ourselves ('You can do it!'), chide ourselves ('Focus!'), and so on. Again, these utterances are intentional actions, performed with the goal of solving our current problem. They are heard and processed like other utterances and interpreted as requiring some response. Type 1 processes get to work on the task and, with luck, generate a further utterance or other action which either solves our problem or takes us a step closer to a solution. Sometimes we conduct a dialogue with ourselves, posing questions and answering them as a way of thinking through the options. We also create extended arguments, moving from one utterance to another in accordance with simple inferential principles we have been taught or have picked up in the course of debate with others. And as with mathematical reasoning, we can combine a variety of techniques to explore a problem space, using utterances as cognitive stepping stones. Language provides an excellent medium for such flexible, reflective thinking, having an open-ended representational capacity and a syntactic structure that facilitates logical inference.

Intentional reasoning can also be done covertly, in the head. Instead of producing overt symbols, sketches, and utterances, we can create mental images of those things. The claim that we can intentionally create mental imagery is not controversial (just try visualizing your front door or saying your address to yourself in inner speech). In the case of inner speech, this probably involves mentally rehearsing the action of saying the words in question (which generates sensory representations of hearing them), but in other cases it seems to involve the intentional direction of attention in order to stimulate sensory activity associated with relevant stimuli or with episodic memories (Carruthers 2015). In either case, the imagery produced has an autostimulatory effect. Attention sustains the representations in working memory, resulting in their being made available ('globally broadcast') to all Type 1 subsystems, which process them as they would representations generated by external stimuli.[4]

Mental imagery allows the internalization of many external problem-solving activities, in particular those using speech. Processes of self-questioning, self-guiding, self-prompting, argument construction, and inner dialogue can

[4] For detailed proposals about the neural mechanisms involved in this kind of sensory-based reflective reasoning, see Carruthers 2006, 2015.

all be conducted silently in one's head.[5] Imagery also allows the development of a wide range of new problem-solving strategies, in which imagined scenarios serve as proxies for aspects of the world. To take an example frequently used in the literature on mental imagery, if you want to know how many windows there are in your house, you can visualize each room in turn and count the windows. Imagery can also be used to evaluate plans and hypotheses before committing to them. If you are trying to decide where to go for a picnic, you can visualize the different candidate locations and see what emotional reactions they evoke. Visual imagery, together with imaged utterances, can thus provide the decoupled 'secondary' representations needed for hypothetical thinking.

This is not the place to attempt a full survey of the various techniques of imagistic autostimulation, but it is safe to say that there are many of them and that they can be flexibly combined in an exploratory way. It is worth stressing that autostimulatory processes needn't be pre-planned. We don't need to know precisely which autostimulations to generate in order to solve a problem. (If we did, then we would in effect already have solved it.) Rather, we follow a process of trial and error and may hit many dead ends before we reach a solution. At the same time, however, the process needn't be completely random. We may have picked up useful tricks and developed hunches about what will work, based on past experience.

Now, my proposal is that the core distinction between Type 1 and Type 2 processing concerns the role of intentional autostimulatory actions. Type 2 processes constitutively involve such actions, whereas Type 1 processes do not. (Since they are not under intentional control, we may continue to speak of Type 1 processes as *autonomous*.) Note that I do not restrict Type 2 processes to ones that occur 'in the head', using sensory imagery. The defining characteristic of Type 2 reasoning is that it involves intentional autostimulatory action. Whether the actions are covert or overt is incidental. Of course, on this view Type 2 processing *also* involves Type 1 processing and is driven by it; but there is still a qualitative difference between the two. Type 1 processes do not involve the performance of intentional actions and are not mediated by perceptions or sensory imagery.

This distinction subsumes the other core distinctions that have been proposed: Intentional autostimulation loads on working memory and supports cognitive decoupling since the perceptual or imagistic representations

[5] Of course, not all intentional reasoning processes can be internalized. When it is necessary to keep referring back to previous steps, as in doing a long division, our working memory capacity is soon exceeded and an external record is required.

involved are held in working memory and can represent non-actual states of affairs. It also explains why Type 2 thinking has the typical correlated features. Autostimulation is conscious because the representations generated are globally broadcast (global broadcast is widely agreed to be sufficient for consciousness in the access sense and at least correlated with consciousness in the phenomenal sense).[6] It is controlled because it is an intentional action, slow and effortful because it requires controlled attention, serial because we can perform only one action at a time, and so on.

5 A virtual mind

This view of Type 2 thinking has implications for the evolution of the new, 'System 2' mind. This did not require the creation of a new general-purpose reasoning system, or indeed of any completely new neural structures. The engine of Type 2 thinking is provided by the collection of specialist perceptual, conceptual, emotional, and motivational subsystems which constitute the old, System 1 mind, and which evolved in response to specific adaptive pressures.[7] The other key ingredients required for Type 2 thinking were almost certainly already in place too. Forms of working memory, attention, episodic memory, and executive control are found in other animals (Carruthers 2015, Ch. 8), and natural language probably developed initially for social purposes.[8]

This suggests that the development of Type 2 thinking was predominantly a process of cultural evolution, involving the discovery and transmission of habits of autostimulation. It is plausible to see this process as the privatization and then internalization of certain social practices. Humans began by cognitively stimulating each other, helping their peers solve problems by offering suggestions, giving advice, asking questions, making sketches, and so on. They also developed practices of public argumentation, setting out arguments in favour of their ideas and plans. Later, they privatized these habits, providing a similar commentary on their own activities and constructing arguments in

[6] In fact, I believe that access consciousness is the only kind there is and that phenomenal consciousness is illusory (Frankish 2016). But that is another — though related — story.

[7] Some writers have argued that humans have a specialist argumentation system (of the Type 1 kind), whose function is to produce rational arguments for use in debate with one's peers (Mercier and Sperber 2011). Such a system would obviously be a great asset in supporting Type 2 thinking, helping to generate cogent arguments in inner speech, but it is still a precursor system, which evolved for social purposes.

[8] Speculating about the origins of language is a notoriously risky business, but I think it is safe to assume that its evolution was initially driven by the needs and opportunities of social life, though its co-option for cognitive purposes may have fostered its further development. I assume that the evolutionary process itself was a combined biological and cultural one (Dennett 2017).

private. Finally, they internalized this commentary and developed further self-stimulatory tricks using mental imagery.

There may have been some relatively minor neural adaptations to support the process. Individuals who had discovered the trick of autostimulation would have had a huge advantage over their peers, creating selectional pressure for neural adaptations that favoured the automatic acquisition and elaboration of the trick—a process known as the Baldwin effect (Dennett 1991). But techniques of intentional reasoning still have to be learned, and a parallel process occurs in child development, as psychologists in the Vygotskyan tradition stress (e.g., Diaz and Berk 1992; Vygotsky 1986; Winsler et al. 2009). Adults *scaffold* children's cognitive development by offering guidance, suggestions, and instructions, which enable children to work through problems they could not have solved on their own. Children then imitate this commentary in self-directed ('private') speech, providing the scaffolding for themselves. Finally, they internalize this private speech as inner speech.

This reinterpretation of dual-process theory casts talk of dual systems in a new light. On this view, there is just one neural system—the collection of 'System 1' subsystems, together with attentional and executive systems and working memory. Note that this claim is compatible with the neuroimaging evidence for dual-*process* theory mentioned earlier. The claim is not that exactly the same subsystems are involved in generating a Type 2 response to a problem as would have been involved in generating a Type 1 response to it. Quite the opposite. Type 2 thinking may bring a different, wider range of neural resources to bear on the problem, and it involves engaging executive and working memory systems as well. The claim is merely that there are no subsystems designed *solely* to support Type 2 thinking.

On this view, then, 'System 2' is not a neural system but a new level of organization, formed by culturally transmitted habits which restructure the activities of the biological brain. In Dennett's phrase, it is a softwired 'virtual machine', like a computer operating system, running on the hardware of the biological brain (Dennett 1991, Ch. 7). If the old mind is a biological mind, then the new mind is a virtual one.

The reader may suspect some sleight of hand here. How could perceptual and imagistic feedback so radically enhance the problem-solving powers of the brain? After all, the knowledge that we draw on in Type 2 thinking is encoded in Type 1 memory systems and available to Type 1 thinking. Why can't Type 1 processes take care of everything? There are several points to make here. First, as Dennett observes, feedback may enable the integration of information from different mental subsystems. Subsystems that lack internal channels of communication can share information by generating speech or sensory imagery expressing it, thereby making it available to perceptual systems and, through them, to the rest of the mind (Dennett 1991). Natural language is ideally suited to this role of content integrator, since most mental subsystems have access to the language system (Carruthers 2006). Second, imagistic feedback is not random but intentionally controlled, directed to solving some specific problem and guided by learned procedures and tricks, as discussed earlier. We learn ways of constructing verbal arguments and exploiting sensory imagery, just as we learn to do maths, drive, or play tennis. Such learning, of course, involves myriad micro-changes to the biological brain, encoding the new beliefs, skills, and habits. Third, Type 2 processing enables us to exploit our existing knowledge in new ways. Our memories encode a vast amount of information, all potentially relevant to any problem we face. Autostimulation has a strong selectional effect. When we ask ourselves a question, many different items of knowledge compete for articulation in inner or outer speech. The ones that win then prime the next round of selection, giving the edge to related items, and so on. Thus, by autostimulating we can hack a path through the informational jungle, making new connections and arriving at new conjectures. Of course, many paths turn out to be dead ends, but with persistence and self-criticism we can find good ones.

To sum up so far: There is robust evidence for a qualitative distinction between two types of thinking, intuitive and reflective. This distinction is best interpreted as one between autonomous processes and intentional reasoning. Autonomous processes can guide everyday behaviour in familiar environments, but intentional reasoning is needed to deal with novel or complex problems. It involves creating overt representations, questioning ourselves, imagining relevant scenes and objects, and constructing arguments in inner speech. The objects and imagery produced act as autostimulations, providing fresh inputs to our autonomous processes, which may then generate a response, in the form of more inner speech, other sensory imagery, or an emotional reaction. This reframes the problem or provides a partial solution to it, and in turn acts as a further autostimulation, and so on. In this way, by engaging in cycles of autostimulation and response, we can work our way through problems that would otherwise be beyond us. Culturally transmitted habits of autostimulation create a new level of mental activity, a virtual mind, which engages in reflective thinking. It is by installing this virtual system in our heads that we come to approximate to general intelligence.

6 Enhancing human intelligence

What implications does this dual-minds view have for the project of artificially enhancing human intelligence? The first thing to ask is which system we are thinking of enhancing: the biological mind or the virtual mind? The methods would need to be very different. Enhancing the biological mind would mean directly interfering with the hardware of the brain. We might seek to boost our cognitive functioning with nootropic drugs, neurostimulation, or genetic manipulation. We might extend our perceptual capacities by hooking up artificial sensors to our sensory cortices, relying on the brain's plasticity to extract the information they supply. More ambitiously, we might create artificial cognitive subsystems, which interface with our biological ones. These would probably have to be self-organizing systems, which could be implanted early in life and grow alongside the biological ones, forming complex low-level connections with them. None of these technologies will be easy to develop, and installing them will require detailed understanding of brain functioning and development.

Enhancing the virtual mind is a completely different matter. Indeed, the virtual mind is itself a cognitive enhancement—a set of tricks for extending the powers of the biological brain, often through the use of artefacts. These tricks created the new human mind, with its powers of hypothetical thinking and creative problem-solving, and it is very tempting to link their emergence with the 'cultural explosion' 30-60,000 years ago, when art, religion, and complex technology first appeared (Mithen 1996). (We might say that the first technological singularity occurred in the Upper Palaeolithic.)

Moreover, the virtual mind itself can easily be enhanced. On a software level, we can learn new reasoning techniques—new procedures for constructing arguments, doing calculations, making decisions, and so on. Much of human education, formal and informal, is concerned with this kind of enhancement. Adding new hardware is easy too. Because we have internalized many tricks of autostimulation, and added new private ones, we tend to think of our conscious minds as essentially private (the streams of consciousness in our heads) and to suppose that enhancing them would require tinkering with our brains. But this is to over-emphasize an incidental feature of intentional reasoning. From a functional perspective, the autostimulatory routines we run in our heads are on a par with public ones involving the manipulation of artefacts, such as writing or sketching. In both cases we intentionally produce and manipulate artefacts and symbols in order to transform complex problems into simpler ones that our biological minds can solve. Technology can vastly extend this process by transforming difficult abstract problems into easy practical ones. Think of using a calculator to solve a complicated mathematical problem. Instead of solving the maths problem itself, we now have to solve the much simpler problem of how to get the calculator to solve it. Again, the process is fundamentally autostimulatory. At each step the calculator provides us with new stimuli, creating new, simpler subproblems: which keys to press first, how to interpret the answer the calculator displays, what entries to key in next, and so on. The solutions to these simpler problems are provided by our Type 1 processes, and the solution to the whole problem is the product of cycles of internal Type 1 processing and external electronic processing, which constitute a temporally and spatially extended Type 2 process.

We also supplement our biological memories with external sources of knowledge, such as tables, reference books, and databases. Rather than posing a question to ourselves, we can consult an external resource, retrieving items of information for use in Type 2 reasoning. Again, from the perspective of the virtual mind there is no significant difference between biological memory and external information sources. Both are resources we intentionally access (by autostimulation in one case, with hands and eyes in the other), in the hope that they will yield reliable and relevant information. External resources merely expand the hardware on which the virtual mind is run.

These enhancements to the virtual mind are easy to install. The devices involved are designed to interface naturally with our biological minds through our hands and sense organs. We press the keys of the calculator and look at its display. So adoption is easy; we just plug in new cognitive aids via sensory interfaces. All that is required is some training in using the devices and interpreting their outputs. (We might be able to make the devices more efficient by developing interfaces that bypass the external organs, detecting motor commands in the brain and sending signals directly to afferent sensory pathways, but such shallow interventions would be relatively easy to accomplish.) For thousands of years, we humans have been enhancing our Type 2 thinking with artefacts, from writing instruments and abacuses through to iPhones and smart glasses, and this sort of enhancement looks set to progress rapidly in coming decades.[9]

[9] For careful exploration of how internet technology is extending and transforming human agency and cognition, and the costs and benefits involved, see Clowes 2017, 2019. Clowes stresses that although current developments have novel features, they continue a long-established process through which the human mind has been re-shaped and enhanced through interactions with material culture.

7 Artificial intelligence

As I noted earlier, from a dual-system perspective, different traditions in AI can be seen as focusing on different mental systems: computational modelling of general intelligence focusing on System 2, and embodied, behaviour-based approaches on System 1. The former project has proved notoriously intractable, and the present view of System 2 sheds some light on this. If System 2 is a virtual system, then in order to reproduce its powers, we would need to reproduce the powers of the biological mind, too—the vast suite of fast, automatic, intelligent subsystems that forms the engine of System 2 thinking. To adopt a top-down approach is to put the cart before the horse, like trying to create an operating system without having the hardware to run it on. While a virtual mind may be easy to enhance, it is difficult to create.

In principle, no doubt, general intelligence could be modelled directly from the top down, perhaps even in computational terms, but it would be a formidable challenge. (If this isn't obvious, consider that it would involve, among other things, finding ways of representing all the diverse kinds of Type 1 knowledge in a format that allows for their integration in reasoning; devising procedures for rapidly retrieving contextually relevant items from a vast knowledge base; and creating a powerful general reasoning system that can perform a wide range of operations, including belief fixation and updating, decision making, planning, causal reasoning, mentalizing, language processing, abductive inference, and creative thinking.) Moreover, it is unclear what the target of the project would be. It is tempting to take our Type 2 thought processes as the paradigm of general intelligence, but we should not idealize them. Human Type 2 thinking is shaped by many contingent factors: by the nature and capacities of the specialist subsystems that drive it, by the cultural resources available for its programming, and by individual differences in the way we conduct it. If we were trying to model general intelligence computationally, it is not obvious that we should focus on our own idiosyncratic, species-specific and culture-specific form of it (unless, of course, we want to create artificial versions of ourselves).

A more practicable approach to creating general intelligence would be to work from the bottom up, creating independent creatures with Type 1 minds and coaxing them into developing Type 2 minds for themselves. We would need to equip them with goals, social instincts, suites of perceptual, cognitive, and motivational systems, and a communication system. By tuning their goals in the right way, we might get them to start cognitively stimulating each other and then autostimulating, working their way gradually toward explicit Type 2 thought. It is unlikely, however, that we could ensure this outcome through engineering alone. Our creatures would need to develop social institutions and cultural practices in order to sustain and transmit the skills and knowledge required for Type 2 thinking. As deliberate designers we could only take the process so far, but as guides and teachers we could take it further, sharing the mental software that has made us who we are. We might train our creatures, as we train children, providing scaffolding that helps them learn how to think. 'What might help?' 'What do you need to know?' 'Could you look at it differently?' 'What if you did this?' Our interactions with AIs may be much like those with precocious children.[10]

It may be, then, that the best way to create general intelligence will be to create beings who can create it for themselves. If so, then AIs will also have two minds, though the shape of both will probably be quite different from ours. The form of Type 2 thinking is determined by the nature of the autostimulatory mechanisms employed (the language system, perceptual and imagistic abilities, working memory capacity, and so on), and the virtual minds of AIs might be much richer and more complex than ours.

8 The risks of enhancement and AI

Speculation about enhanced and artificial intelligence soon turns to concerns about the risks involved, and I shall close this chapter with some remarks on this from a dual-minds perspective.

A common worry is that, having embarked on the creation of artificial intelligence, we may lose control of the process. Our creations may take control of their own development, pursue their own projects, and become indifferent or hostile to us. I think this is alarmist. For AIs to take control in this way, they would need to be capable of flexible, creative thinking of the Type 2 kind. They would need to be able to set themselves new goals, evaluate hypothetical scenarios, plan ahead, and much more. But, as we have seen, such abilities won't be easy to engineer, and a more feasible strategy will be to create artificial creatures with animal-like intelligence, and then help them to bootstrap themselves into general intelligence though cultural processes. This is unlikely to be a fast or straightforward process. We worry about AIs developing rapidly and escaping our control, but it is more likely that we shall have to nurture them laboriously through a long childhood, both as an artificial species and as individuals. Before we have to deal with super-intelligent AI overlords, we shall probably have to spend many years dealing with demanding, reckless,

[10] For a related perspective, which explores the role of language use in developing a variety of higher cognitive functions in robotic systems, see Mirolli and Parisi 2011.

accident-prone, and occasionally brilliant artificial children.

There is, however, another way in which we may cede control to technology, which is a much more pressing concern. I stressed how easy it is to enhance our virtual minds, using artefacts to transform problems and to supplement our biological memories. We have been enhancing and extending our virtual minds in this way for thousands of years, and our modern minds are heavily dependent on external support. (Think what effect the loss of your phone would have on your ability to do your job or organize your life.) Modern technology is accelerating this process, however, offering increasingly powerful new cognitive aids. Programming our biological brains to support Type 2 thinking is a laborious business, which involves mastering complex reasoning procedures and memorizing vast amounts of information. Computer technology offers shortcuts. Instead of learning to do long division, we can learn to use a calculator; instead of memorizing historical facts, we can learn to access an online encyclopaedia; instead of memorizing spellings, we can learn to run a spellcheck program. Technology looks set to supply us with ever more powerful shortcuts like this, allowing us to offload cognitive drudgery onto electronics in the way that previous generations offloaded manual labour onto mechanical appliances.

We can also expect technology to give us many completely new capacities, supplementing our biological minds with external modules, tightly linked via sensory interfaces. We shall be able to query these modules for information, entertainment, and motivational stimuli, and use them to make visual, aural, and tactile contact with far-off people and places. We can expect our conscious minds to be radically enriched, allowing us to develop new ways of working, socializing, and loving.

The advantages of all this are obvious, and we shall probably find them impossible to resist. (Why should a lawyer spend years studying case law if they can buy a tiny earpiece that will instantly retrieve contextually relevant data as needed and feed it to them?) But the dangers are obvious too. Making our conscious minds dependent on external electronic hardware as well as our biological brains will be a risky business. Our brains are robust, well-protected organs, which are the product of millions of years of natural R&D and have a remarkable capacity for self-repair. Electronic devices are far more vulnerable. A solar flare might knock them out and leave us cognitively disabled. And if they fail, it won’t be easy to fall back on older technology. (Who now knows how to use a slide rule?)

More worryingly perhaps, we shall be at the mercy of those who control the technology. Having offloaded so much of our skill and knowledge, we won’t have the resources to assess the value of the information and guidance we are fed, and those who control the feed will be able to manipulate the rest of us. We are already seeing something like this in the use of social media bots to manipulate opinion during elections. Seemingly relevant images and bits of information pop up on social media, just as thoughts pop into our heads, and it is easy to let them guide one's thoughts and decisions. Imagine having a host of similar bots whispering in your ear, guiding your work, your social relations, your personal life, your very thinking.

The moral, then, is that it is not the master AIs we should worry about but the servant ones. We may end up developing our virtual minds to the point where they are no longer really ours, no longer tethered to our biological minds and to the purposes and values rooted there. This is the paradox of the virtual mind. In learning how to manipulate our biological minds and create virtual minds for ourselves, we risk undermining the locus of purpose and control that our biological minds sustained. It is the price of being creatures with two minds.[11]

References

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159. doi:10.1016/0004-3702(91)90053-M

Carruthers, P. (2006). *The architecture of the mind: Massive modularity and the flexibility of thought*. Oxford: Oxford University Press.

Carruthers, P. (2009). An architecture for dual reasoning. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 109–127). Oxford: Oxford University Press.

Carruthers, P. (2015). *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford: Oxford University Press.

Chaiken, S. & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Chen, S. & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York: Guilford Press.

Clowes, R. W. (2017). Extended memory. In S. Bernecker & K. Michaelian (Eds.), *The Routledge handbook of philosophy of memory* (pp. 243–254). Abingdon: Routledge.

Clowes, R. W. (2019). Immaterial engagement: Human agency and the

[11] I am grateful to Wim De Neys and two anonymous reviewers for comments on earlier drafts of this chapter. Some of the ideas in the chapter were first sketched in an interview for *Interalia* magazine (https://www.interaliamag.org/interviews/keith-frankish/), from which a few sentences are drawn.

cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences*, 18(1), 259–279. https://doi.org/10.1007/s11097-018-9560-4

De Neys, W. (Ed.). (2018). *Dual process theory 2.0*. New York: Routledge.

Dennett, D. C. (1991). *Consciousness explained*. New York: Little, Brown and Co.

Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. New York: W.W. Norton & Company.

Descartes, R. (1984). *The philosophical writings of Descartes: Volume 2*. (Trans., J. Cottingham, R. Stoothoff, & D. Murdoch). Cambridge: Cambridge University Press.

Diaz, R. M. & Berk, L. E. (Eds.). (1992). *Private speech: From social interaction to self-regulation*. Hillsdale, NJ: Lawrence Erlbaum.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *The American psychologist*, *49*(8), 709–724.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove: Lawrence Elrbaum Associates.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove: Psychology Press.

Evans, J. St. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford: Oxford University Press.

Evans, J. St. B. T. & Over, D. E. (1996). *Rationality and reasoning*. Hove: Psychology Press.

Evans, J. St. B. T. & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. doi:10.1177/1745691612460685

Frankish, K. (1998). Natural language and virtual belief. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 248–269). Cambridge: Cambridge University Press.

Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.

Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 89–107). Oxford: Oxford University Press.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, *5*(10), 914–926. doi:10.1111/j.1747-9991.2010.00330.x

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, *23*(11–12), 11–39.

Frankish, K. (2018). Inner speech and outer thought. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices* (pp. 221–243). Oxford: Oxford University Press.

Frankish, K. & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–29). Oxford: Oxford University Press.

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology*, *20*(6), 733–743. doi:10.1177/0959354310378184

Gleick, J. (1992). *Genius: The life and science of Richard Feynman*. New York: Pantheon Books.

Heyes, C. M. (2018). *Cognitive gadgets: The cultural evolution of thinking.* Cambridge, MA: Harvard University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Keren, G. & Schul, Y. (2009). Two is not always better than one. *Perspectives on Psychological Science*, *4*(6), 533–550. doi:10.1111/j.1745-6924.2009.01164.x

Kruglanski, A. W. & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. doi:10.1037/a0020762

Melnikoff, D. E. & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, *22*(4), 280–293. doi:10.1016/j.tics.2018.02.001

Mercier, H. & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(02), 57–74. doi:10.1017/S0140525X10000968

Mirolli, M. & Parisi, D. (2011). Towards a Vygotskyan cognitive robotics: The role of language as a cognitive tool. *New Ideas in Psychology*, *29*(3), 298–311. doi:10.1016/j.newideapsych.2009.07.001

Mithen, S. J. (1996). *The prehistory of the mind: A search for the origins of art, religion and science*. London: Thames & Hudson.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988–1010.

Osman, M. (2018). Persistent maladies: The case of two-mind syndrome. *Trends in Cognitive Sciences*, *22*(4), 276–277. doi:10.1016/j.tics.2018.02.005

Pennycook, G., Neys, W. D., Evans, J. St. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The mythical dual-process typology. *Trends in Cognitive Sciences*, *22*(8), 667–668. https://doi.org/10.1016/j.tics.2018.04.008

Petty, R. E. & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in experimental social psychology*, *19*, 123–205.

Sloman, S. A. (1996). The empirical case for two systems of reasoning.

Psychological Bulletin, *119*(1), 3–22.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.

Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford: Oxford University Press.

Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought*. New Haven: Yale University Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.

Stanovich, K. E. & Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, *11*(1), 3–13. doi:10.1007/s11299-011-0093-6.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(05), 645–726.

Steels, L. & Brooks, R. A. (Eds.). (1995). *The artificial life route to artificial intelligence: Building embodied, situated agents*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Vygotsky, L. (1986). *Thought and language*. (Trans. and Ed., A. Kozulin). Cambridge, MA: MIT Press.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Winsler, A., Fernyhough, C., & Montero, I. (Eds.). (2009). *Private speech, executive functioning, and the development of verbal self-regulation*. Cambridge: Cambridge University Press.