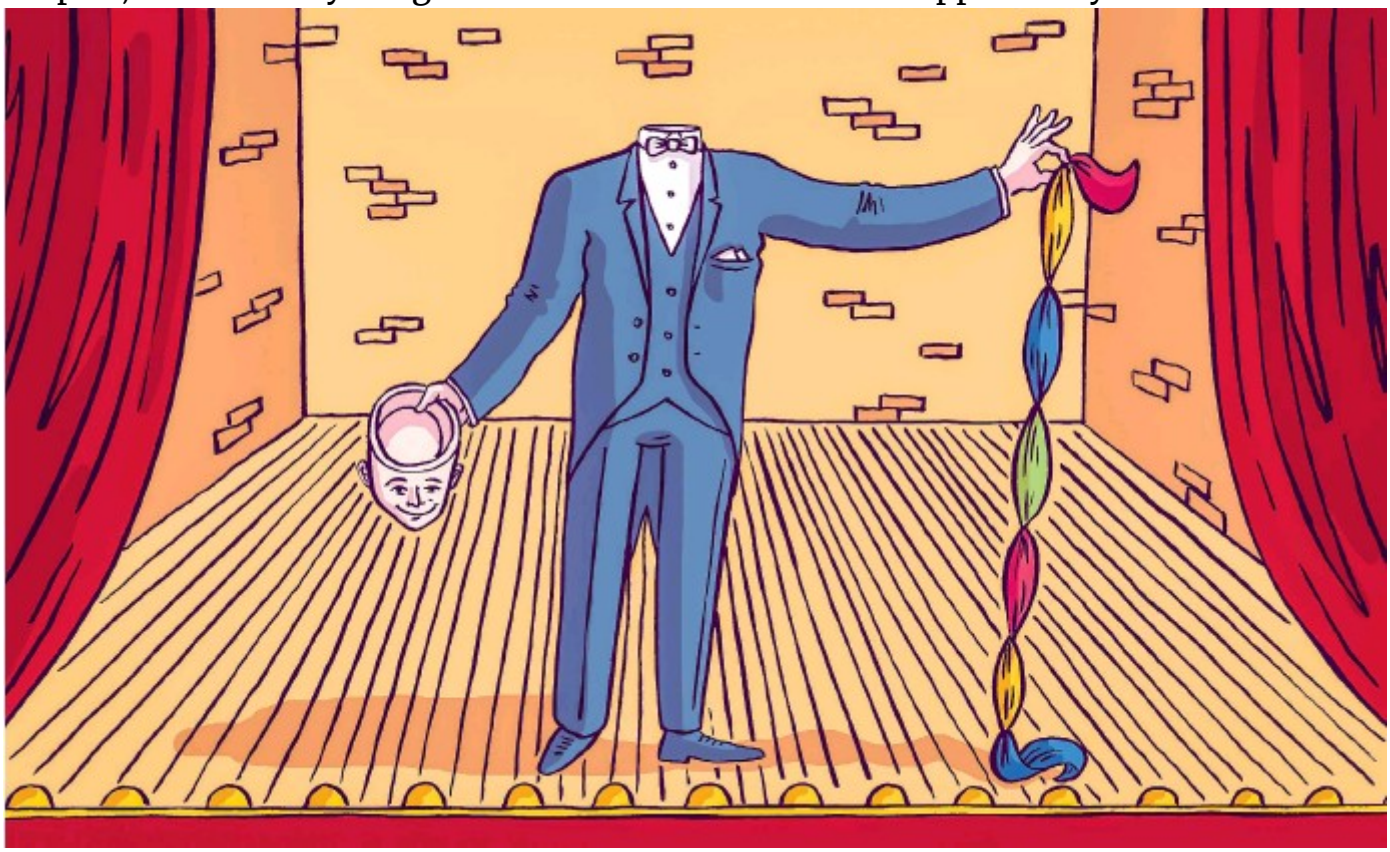
Time for a revolution in our sense of self?

In a radical reassessment of how the mind works, a leading behavioural scientist argues the idea of a deep inner life is an illusion. This is cause for celebration, he says, not despair

The Observer Domestic edition · 1 apr. 2018 · Nick Chater

At the climax of Anna Karenina, the heroine throws herself under a train as it moves out of a station on the edge of Moscow. But did she really want to die? Had the ennui of Russian aristocratic life and the fear of losing her lover, Vronsky, become so intolerable that death seemed the only escape? Or was her final act mere capriciousness, a theatrical gesture of despair, not seriously imagined even moments before the opportunity arose?



We ask such questions, but can they possibly have answers? If Tolstoy says that Anna has dark hair, then Anna has dark hair. But if Tolstoy doesn't tell us why Anna jumped to her death, then Anna's motives are surely a void. We can attempt to fill this void with our own interpretations and debate their plausibility. But there is no hidden truth about what Anna really wanted, because, of course, Anna is a fictional character.

Suppose instead that Anna were a historical figure and Tolstoy's masterpiece a journalistic reconstruction. Now Anna's motivation becomes a matter of history, rather than a literary interpretation. Yet our method of inquiry remains the same: the very same text would now be viewed as providing (perhaps unreliable) clues about the mental state of a

real person, not a fictional character. Historians, rather than literary scholars, might debate competing interpretations.

Now imagine that we could ask Anna herself. Suppose the great train slammed on its brakes just in time. Anna, apparently mortally injured, is conveyed in anonymity to a Moscow hospital and, against the odds, pulls through. We catch up with Anna convalescing in a Swiss sanatorium. But, as likely as not, Anna will be as unsure as anyone else about her true motivations. After all, she too has to engage in a process of interpretation as she attempts to account for her behaviour. To be sure, she may have "data" unavailable to an outsider – she may, for example, remember the despairing

words "Vronsky has left me forever" running through her mind as she approached the edge of the platform. However, any such advantage may be more than outweighed by the distorting lens of self-perception. In truth, autobiography always deserves a measure of scepticism.

There are two opposing conclusions that one might draw from this vignette. One is that our minds have dark and unfathomable "hidden depths". From this viewpoint, we cannot expect people to look reliably within themselves and compile a complete and true account of their beliefs and motives. Psychologists, psychiatrists and neuroscientists have long debated how best to plumb the deep waters of human motivation. Word associations, the interpretation of dreams, hours of intensive psychotherapy, behavioural experiments, physiological recordings and brain imaging have been popular options.

I believe, though, that our reflections should lead us to a different conclusion: that the interpretation of real people is no different from the interpretation of fictional characters. If Tolstoy's novel had been reportage, and Anna a living, breathing member of the 19th-century Russian aristocracy, then, of course, there would be a truth about whether Anna was born on a Tuesday. But, I argue, there would still be no truths about the real Anna's motives. No amount of therapy, dream analysis, word association, experiment or brain scanning can recover a person's "true motives", not because they are difficult to find, but because there is nothing to find.

Evidence of a hoax

This is not a conclusion I have come to lightly. As a psychologist, I want to understand how people think and decide. It would be awfully convenient if the rich stories we tell about our own thoughts were at least roughly on the right track; if they just needed to be tidied, pruned and generally knocked into shape to get a true picture. It would be convenient, but utterly wrong. The weight of evidence against the reality of "mental depth" is simply overwhelming. Having resisted the evidence for years, I've finally admitted defeat.

Perception provides some ominous clues. Consider Jacques Ninio's wonderful "12 dots" illusion. Twelve black dots are arranged in three rows of four dots each. The dots are large enough to be seen clearly and simultaneously against a white background. But when arranged on the grid, they seem only to appear when you are paying attention to them. Dots we are not attending to are somehow "swallowed up" into the diagonal grey lines. Interestingly, we can pay attention to adjacent pairs of dots, to lines of dots, to triangles and even squares – although these are highly unstable. But our attention is in short supply and, where we are not attending, the dots disappear.

Remarkably, the limits of attention apply just as well as we scan our everyday environment: we can attend to just one object at a time – the other objects are effectively invisible. Our sense that we can grasp the entire visual world in full detail and colour is, then, a hoax. Instead, we see through a remarkably narrow "window" of attention, grasping just one object, word or face at a time. But the hoax is sustained because, as soon as we wonder about, say, the colour of a vase or the identity of a word, our eyes and our attention can, almost instantly, flick into action, lock on the "target" and answer our question. And the answer is created so fluently that we imagine that it was there all along.

By extension, then, we may begin to doubt our phenomenology of a rich inner world, teeming with ideas and feelings. Indeed, it turns out that here, too, our brains have been inventing wildly. To pick one particularly striking example, let us consider the remarkable classic studies of cognitive neuroscientist Michael Gazzaniga on patients with "split brains", whose left and right brain hemispheres have been surgically severed.

Our brains have "crossover" wiring: the left hemisphere sees the right half of the visual world and controls the right hand, and vice versa. So this means that, for split-brain patients, the right and left hemispheres can be shown entirely different stimuli and make wholly independent responses. In a famous demonstration, Gazzaniga shows a snowy scene to the right hemisphere and a chicken's foot to the left. The right hemisphere has to find a picture that matches what it sees (the snowy scene) and naturally enough chooses a picture of a shovel (with the left hand). How does the left hemisphere (the seat of language) explain this choice? It should be baffled, because it knows nothing about the real cause of the right hemisphere's choice, because it can't see the snowy scene. Yet, quick as a flash, it has a ready answer: the chicken's foot is associated with a chicken and you need a shovel to clean out the chicken shed. Elegant, but entirely wrong.

Our language system is continually generating a flow of plausible-sounding explanations of the reasons behind our actions but, suspiciously, the flow continues with the same speed and confidence when our language system cannot possibly know the truth. And it continues without balking. It was confabulating all along.

Our inner, mental world is a work of the imagination. We invent interpretations of ourselves and other people in the flow of experience, just as we conjure up those of fictional characters from a flow of written text. Returning to Anna, we can wonder whether she despaired primarily of her precipitous social fall, the future of her son or the meaninglessness of aristocratic life, rather than being tormented by love. There is no ground truth about the right interpretation, though some are more compelling and better evidenced in Tolstoy's text than others. But Tolstoy, the journalist, would have nothing more than interpretations of the "real" Anna's behaviour; she could only venture one more interpretation of her own behaviour.

The unfolding of a life is not so different to that of a novel. We generate our beliefs, values and actions in the moment. Thoughts, like fiction, come into existence in the instant that they are invented and not a moment before. The sense that behaviour is merely the surface of a vast sea, immeasurably deep and teeming with inner motives, beliefs and desires is a conjuring trick played by our own minds. The truth is not that the depths are empty, or even shallow, but that the mind is flat: the surface is all there is.

The improvised mind has an answer for everything. Each choice, preference or belief small and large can, when challenged, yield an easy flow of rationalisation. Why this sofa? Why Bach, not Brahms? Why this choice of career? Why children or not? Why evolution, not creationism? How does a bicycle work, or a violin, or a currency? And each justification can be buttressed with further justifications, caveats and clarifications, and each of these be defended further, seemingly without end. Our creative powers are so great, and so effortless, that we can fancy we must be consulting an "inner oracle", which can look up preformed answers to each question.

One crucial clue that the inner oracle is an illusion comes, on closer analysis, from the fact that our explanations are less than watertight. Indeed, they are systematically and spectacularly leaky. Now it is hardly controversial that our thoughts seem fragmentary and contradictory. I can't quite tell you how a fridge works or how electricity flows around the house. I continually fall into confusion and contradiction when struggling to explain rules of English grammar, how quantitative easing works or the difference between a fruit and a vegetable.

But can't the gaps be filled in and the contradictions somehow resolved? The only way to find out is to try. And try we have. Two thousand years of philosophy have been devoted to the problem of "clarifying" many of our commonsense ideas: causality, the good, space, time, knowledge, mind and many more; clarity has, needless to say, not been achieved. Moreover, science and mathematics began with our commonsense ideas, but ended up having to distort them so drastically – whether discussing heat, weight, force, energy and many more – that they were refashioned into entirely new, sophisticated concepts, with often counterintuitive consequences. This is one reason why "real" physics took centuries to discover and presents a fresh challenge to each generation of students.

Philosophers and scientists have found that beliefs, desires and similar every-day psychological concepts turn out to be especially puzzling and confused. We project them liberally: we say that ants "know" where the food is and "want" to bring it back to the nest; cows "believe" it is about rain; Tamagotchis "want" to be fed; autocomplete "thinks" I meant to type gristle when I really wanted grist. We project beliefs and desires just as wildly on ourselves and others; since Freud, we even create multiple inner selves (id, ego, superego), each with its own motives and agendas. But such rationalisations are never more than convenient fictions. Indeed, psychoanalysis is projection at its apogee: stories of greatest possible complexity can be spun from the barest fragments of behaviours or snippets of dreams.

An experiment in artificial intelligence

Yet perhaps our thoughts and actions may be guided by "commonsense theories" that, though different from scientific theories, could be coherent nonetheless. This is a seductive idea. Starting in the 1950s, decades of intellectual effort were poured into a particularly sophisticated and concerted attempt to crystallise some of our commonsense theories. The goal was to systematise and organise human thought to replicate it and create machines that think like people.

Early attempts to create artificial intelligence followed this approach. Hopes were high. Over successive decades, leading researchers forecast that human-level intelligence would

be achieved within 20 to 30 years. By the 1970s, serious doubts began to set in. By the 1980s, the programme of mining and systematising knowledge started to grind to a halt. Indeed, the project of

coaxing the "theories" from our inner oracle failed in a particularly instructive way. Drawing out the knowledge, beliefs, motives and so on that underpinned people's behaviour turned out to be hopelessly difficult.

Chess grandmasters, it turns out, can't really explain how they play chess, doctors can't explain how they diagnose patients and none of us can remotely explain how we understand the everyday world of people and objects. What we say sounds like explanation – but really it is a barely coherent jumble. Perhaps the single

Inventing our future selves

most important discovery from the first decades of artificial intelligence is just how profound and irremediable this problem is.

The project of modelling artificial on human intelligence has since been quietly abandoned. Instead, over recent decades, AI researchers have made advances by building machines that learn not from people but from direct confrontation with huge quantities of data: images, speech waves, linguistic corpora, chess games and so on. Much of AI has m... tated into a distinct but related field: machine learning. This has been possible because of advances on a number of fronts: computers have become faster, data-sets larger and learning methods cleverer. But at no stage have human beliefs been mined or commonsense theories reconstructed.

The spectacular improvisation of the human mind is, I believe, the core of human intelligence and the ability that allows us to deal so successfully with the complex, open-ended challenges thrown at us by our physical environment and the social world. AI and robotics have succeeded precisely where those improvisational abilities are not required: in the pristine worlds of chess, Go and car assembly plants, for example. Don't be fooled: the "rise of the robots" is no more than supersophisticated automation. The amazing creativity of your brain, as it helps us improvise our way through daily life, won't be replicated in silicon in the near future, perhaps never. Don't despair. This does not mean there isn't something we can define as a "self". Our brains are relentless and compelling improvisers, creating the mind, moment by moment. But, as with any improvisation, in dance, music or storytelling, each fresh thought is not created out of nothing, but built from the fragments of past improvisations. So each of us is a unique history, together with a wonderfully creative machine for redeploying that history to create new perceptions, thoughts, emotions and stories. The layering of that history makes some patterns of thought natural for us, others awkward or uncomfortable. While drawing on our past, we are continually reinventing ourselves, and by directing that reinvention, we can shape who we are and who we will become.

So we are not driven by hidden, inexorable forces from a dark and subterranean mental world. Instead, our thoughts and actions are transformations of past thoughts and actions and we often have considerable latitude, a certain judicial discretion, regarding which precedents we consider, which transformations we allow. As today's thought or action are tomorrow's precedents, we are reshaping ourselves, moment by moment.

This viewpoint contradicts the Freudian inner depths, but it meshes naturally with the cognitive behavioural therapy (CBT) for which there is the best clinical backing. Reshaping our thoughts and actions is hard and requires establishing new patterns of thought and behaviour that overwrite the old – opening up productive channels along which our thoughts may more happily and productively flow. CBT aims to do precisely this: to establish new behaviours (to approach, rather than avoid, a phobic object) and thoughts (shifting thoughts away from negative ruminations) and to create new precedents that may, slowly, come to dominate the old. Therapies of all kinds can help us rewrite the story of our past, to create traditions of thought and action that more constructively address the future. What therapy does not, and cannot do, is to reveal pathologies lurking in our innermost depths – not because those depths are murky, but because they are nonexistent.

This is all very well, you may say. But surely we need beliefs and motives to explain why our thoughts and behaviour make sense, rather than being a completely incoherent jumble. Surely there are crucial inner facts about us, large and small, that set the course of our actions: the things we value, the ideals we believe in, the passions that move us. But if the mind is flat, despite the stories we tell about ourselves and each other, beliefs and motives cannot be driving our behaviour – because they are a projection rather than a reality.

But layers of precedents – the successive adaptation and transformation of previous thoughts and actions to create new thoughts and actions – can provide a different, and more compelling, explanation for the orderly (and, on occasions, the disorderly) nature of thought. In particular, our culture can be viewed as a shared canon of precedents – things we do, want, say, or think – that create order in society as well as within each individual.

By laying down new precedents, we incrementally and collectively create our culture, but our new precedents are based on old, shared precedents, so that our culture also creates us. Considered in isolation, our "selves" turn out to be partial, fragmentary and alarmingly fragile; we are only the most lightly sketched of literary creations. Yet, collectively, we can construct lives, organisations and societies, which can be remarkably stable and coherent.

This is, I believe, a liberating thought. We are not driven by hidden motives, bound by unconscious forces or hopelessly imprisoned by our past. Each new thought and action is a chance to reshape ourselves, if only slightly. Our freedom has its limits, of course. Amateur saxophonists can't "freely" choose to play like Charlie Parker, new learners of English can't spontaneously emulate Sylvia Plath and physics students can't spontaneously reason like Albert Einstein.

New actions, skills and thoughts require building a rich, deep mental tradition; there is no shortcut to the thousands of hours needed to lay down the traces on which expertise is based. Each of us is a unique tradition from which our new thoughts and actions are created. So each of us will play music, write and think in our own way. Yet the same points arise in our everyday lives, our fears and worries, our sometimes bumpy interactions with other people. Our freedom consists not in the ability to transform ourselves magically in a single jump, but to "reshape" our thoughts and behaviours, one step at a time. Our current thoughts and actions are continually, if slowly, reprogramming our minds.

Does this viewpoint imply that we are "blank slates", on which any mental patterns can be written? Not at all. Musical traditions build on the rhythmic pattern generators in our nervous systems, the way our brain groups sounds as voices and much more. Linguistic traditions are shaped by our vocal apparatus, how our brains generate and recognise complex sequences and so on. Human music and language can take many forms – but not any form. Traditions of thought are no different; they, too, will be profoundly shaped by the biases and predilections of our brains – and our genes.

So our thoughts and behaviour are influenced by, but not determined by, biology; and neither are we hemmed in by occult psychic forces within us. Any prisons of thought are of our own invention and can be dismantled just as they have been constructed. If the mind is flat – if we imagine our minds, lives and culture – we have the power to imagine an inspiring future and to make it real.

As today's thought or action are tomorrow's precedents, we are reshaping ourselves, moment by moment