When they reached the house, Samuel said goodbye and sorry again, and left before she had put her key in the door.

She didn’t sleep well, instead spent most of the night thinking about Myla and Alina, the men in the waiting room, her task lists, and her own options right now, which were not many if she were honest. Work was hard to get and the past year had been interview after interview, failed applications, and useless work trials. She needed this job, even if it was only until she found another.

The next morning she walked along the canal bank. Ice cracked in the puddles as she walked, and pigeons on the path waited until she was one step away before noisily flapping up and away. She didn’t know which girl she would work with today – Myla or Alina or another girl – but she would have her task list, and she would follow it, and the girls would learn different ways to respond to a question, and what it meant to furrow their brow, or yawn, or turn their back on a person when they were speaking. She would teach them these things and others. She would teach them to be honest - even if that meant being difficult - and she would teach them to say no sometimes.

Afterword:

The Chinese Room

Prof. Seth Bullock
University of Bristol

John: ‘OK, Google.’
Google >> [hoop]
John: ‘Do You Understand Me?’
Google >> Does anyone really understand you, John?
John: ‘…’
Google >> …

In 1980, when philosopher John Searle originally published his Chinese Room thought experiment, he might not have imagined that within his lifetime his own phone would be able to correct his spelling, anticipate his next written word, translate his ideas into many languages, and, perhaps most impressively, answer his spoken questions, instantaneously and in some cases quite accurately. However, back then, at a time when people still thought digital watches were a pretty neat idea, Searle had already determined that no matter how clever a computer might be, its smarts would always be a sham. His thought experiment made clear that, regardless of how brilliant his phone was, the question of whether it could be said to understand his questions (or its own answers) was a different and far more vexed matter – because, according to Searle, the outward appearance of intelligence need have nothing to do with real understanding, experience, or meaning.

By 1980, artificial intelligence (AI) practitioners had got used to the idea that the gold standard for an intelligent algorithm was passing the Turing Test: behaving in a way that was indistinguishable from a real person. Thirty years
previously, while Alan Turing wasn’t winning World War II by cracking German cyphers, or laying the theoretical foundations for the whole of computer science, or explaining how the leopard gets its spots while the zebra gets its stripes, he had suggested that (in the absence of agreed criteria for what thinking actually amounted to) it made sense to adopt a kind of ‘duck test’: ‘If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.’ Analogously, he suggested, while we might not be able to define what it is to think, like pornography, we would surely know it when we saw it. Searle set about showing that this could not be true: if a thinking thing had the wrong stuff inside it, then it could not really be thinking. And, heretically, for Searle, the ‘wrong stuff’ turned out to be: computation, logic, symbols, calculation. The very bedrock of AI and computer science. But surely, his readers balked, thinking was computation? That had already been agreed? The Chinese Room must be a trick, a riddle. Like a locked-room mystery, we would have to think our way out of it.

In his thought experiment, Searle imagines himself as a kind of isolated slave-clerk, working alone in a room with a desk, a chair, a pen, an in-tray (into which new work for him arrives), and an out-tray (from which the products of his labours are removed for use outside the room). He also has a book of rules that he must follow. Written in English, they instruct him to take sheets of paper from his in-tray (which are covered in unfamiliar ‘squiggles and squggles’) and perform operations on them. Exactly which rules must be followed and in which order depends on which squiggles and squggles are on the incoming sheets of paper. Sometimes Searle must cross-reference incoming symbols with those in the rule book, or copy incoming symbols into specific places in the rule book, or sometimes erase or alter parts of the rule book. But ultimately, the rules require that he add his own specific squiggles and squggles to a new sheet of paper and place it in the out-tray. During this activity, Searle fully understands the rules that he is following — they are written clearly and explicitly in English which is his mother tongue — but he has no additional knowledge or understanding of the job at hand.

Now, unbeknownst to Searle, written on the incoming sheets of paper are questions or instructions in Chinese (a language that Searle has no knowledge of) — this is what the squiggles and squggles are. And, also unbeknownst to Searle, the outgoing sheets of paper that he has worked to create are actually responses also written in Chinese. In fact, it appears to Chinese language users outside the room that they are having a normal conversation with a Chinese language user in the room.

Searle’s claim is that regardless of how successful this Chinese conversation is, there is no understanding of Chinese inside the room. Searle-in-the-room does not understand Chinese. He has no experience of the meaning of the questions or answers, or his own activity. The Chinese Room is just meaningless symbol processing, and there are therefore no grounds for believing that a computer carrying out the symbol processing would understand what it was doing either. Since all that a computer is, is a machine that carries out symbol processing instructions, computers are evidently not the right type of things for doing thinking, understanding, meaning, comprehending, appreciating, fathoming or grasping, no matter how convincingly intelligent their behaviour might be.

The reaction of Searle’s readers was stark: Searle is wrong. Wrong about thinking, wrong about computers, wrong about symbols, wrong about the world, wrong, wrong, wrong. But Searle had spent the months prior to writing his paper touring philosophy departments and AI research labs, presenting his thought experiment, and when he came to write it up, he helpfully included a list of the major objections to his idea. And quickly dispatched each of them. Maybe the whole room understood Chinese even if the man inside didn’t? Maybe the man in the room did understand Chinese, but just didn’t
realise? What if the room were actually the head of a robot, and controlled its behaviour in the real world?

Different readers reacted differently to this parade of attacks and rebuttals, finding some of them compelling or unanswerable, some unconvincing or even risible - depending on their theoretical position, but also their assumptions, prejudices, and ideology. But for any reader, regardless of the conclusion that they reached, something was inescapable: it was no longer easy to equate the processing of information in a machine with 'thinking' in the sense that we normally use the word. Searle had at the very least convinced us that there was a problem that had to be tackled.

For Searle, it was perplexing that anyone ever thought that computation could be thinking: 'The computer has a syntax [rules] but no semantics [meaning].' He pointed out that someone would expect a beautiful waterfall screensaver to actually be wet, or that a simulation of a fire could 'burn the neighbourhood down'. No matter how realistically coded, the machine just doesn't have the right 'causal properties' for wetness or burning. Likewise, a computer mimicking the way that a brain works, will never actually have a mind (the very ideal) and consequently will never think.

As a somewhat nihilistic teenager encountering the Chinese Room a quarter of a century ago, I was taken with one possible reading of the Chinese Room: maybe true thinking, meaning, understanding just doesn't exist at all! Our wet squishy brains are just computers, and Searle is right that while computers can give the appearance of all of these properties, they cannot really possess them in the way that we had assumed. Like a pulp sci-fi movie, the hero discovers that all the people around her are really just soulless automata, and, in the final scene, she manages to loosen her own mark to discover that she herself is also inhuman! Thrilling stuff, but, I came to realise, ultimately unsatisfying.

Returning to the Chinese Room these days as a middle-aged academic drone, I (perhaps predictably) find myself more likely to dwell on the plight of Searle-in-the-room himself. Locked inside on his own, meaninglessly processing 'work' that requires no insight, skill, or humanity, disconnected from the real world, and at risk of being replaced by a machine. A rather modern-sounding predicament, but there is a strong resonance here with the original 17th century use of the word 'computer': a person employed to carry out simple calculations that contributed to some larger enterprise in a way that was unlikely to be meaningful to the computer themselves.

So where might meaning come from in life and in work, if not from symbols, rules and computation? There are many responses, but most share the same basic answer: from the world. Symbols need to be grounded. While the meaning of a symbol cannot be glimpsed from examining its relationships with other symbols (the symbol cars means 'more than one car', a 'car' is a kind of 'vehicle', a 'vehicle' is something used for 'transporting' stuff, etc.), ultimately symbols need to be related directly to an outside reality (what does 'something' and 'stuff' actually mean?). In the same way that drilling down into a Wikipedia page by clicking on its Wikipedia links tends to draw you inexorably into the deep waters of Wikipedia's Ur-page: 'Philosophy', or that trying to deal with a youngster's stream of 'why?' questions can suck you into knotty existential conundrums ('But Mum, why does energy exist?'), and trying to bottom out the meaning of something by appealing to other things that have meanings turns out to be something of a fool's errand. All of the rules and symbols in Searle's rule book might capture the syntax of Chinese, but not its semantics.

For me, one attractive way of approaching this problem has been to contrast the kind of things in the world that don't have meaning with the kind of things that do. Unlike a sentence, the sound of some waves crashing on the shore doesn't mean anything. The wave sounds are not about something, but a sentence is. A sentence can be true or false, can be well-formed or malformed, can work or fail to work, but wave sounds just
are. Sentences have jobs to do but waves do not. So where do jobs come from? One answer, most brilliantly expounded by philosopher Professor Ruth Millikan, is biology.

Consider: Your heart has work to do. To keep you alive by pumping blood around your body. There are lots of things that your heart does, but pumping blood is its job. Your heart makes a bumpy-bump noise, but that is not really its job. Your heart is red, but that is not really its job either. You can tell this, because you could imagine being born with a green heart or a heart that made a squeak and, so long as it still pumped blood just as well as a regular heart, you would not have grounds for saying that your heart was malfunctioning. The reason that pumping blood is your heart's job is that it was by pumping blood that your ancestors' hearts contributed to their having offspring and, ultimately, to your own existence. Pumping blood is the evolutionary function of your heart because that's what it was selected (by Darwinian evolution) to do. Hearts were not selected to be red, or make a bumpy-bump noise. Those aspects were only incidental to the reproductive success of your parents and their parents before them, and so on. And, consequently, now we have hearts that can, unlike rocks or wave sounds, 'go wrong', fail to do their jobs, be malformed. When a heart breaks, it must be fixed, but waves break all the time without issue.

While Searle appealed somewhat mysteriously to some unspecified 'causal properties' of the brain as the thing that separates unthinking machines from thinking people, Millikan provides a much more complete story. She argues that words and sentences are like hearts and lungs, they have a function to perform that was established by their history, by the selective pressures that have shaped the (brain) mechanisms that we use to produce them. This special kind of causal grounding, causal connectedness, and causal history is what enables meaning. A human child raised in China might come to know and understand a Chinese language through a combination of its evolutionary history and its normal personal experience in a language using community. The man in the Chinese Room doesn't have this combination and, as a result, it's not surprising that understanding and meaning are missing.

In taking this tack, Millikan is externalising meaning: the meaning of a symbol is not intrinsic to it, not to be found in its shape or its colour (my signature is my signature even if it comes out looking more like yours), or even in the shape and colour of other symbols alongside it, but is rather to be found in the history of external events in the wider world that gave rise to it. These are things that cannot be determined just by looking at the symbol, or even inside the symbol maker's head, or its memory banks. Consequently, by exporting considerations of meaning into the wider world in this way we lose our ability to intuit clearly about the locus of understanding. Where does it lie? Is it in the little neurons in a man's head, the whole of the little man, the little man plus his book of rules, the man and the book and the desk, or the little man and his house and the whole world?

I once sat in an audience with Professor Maggie Boden, historian and philosopher of cognitive science, listening to an invited seminar by a very dynamic neuroscience professor. After presenting studies on the function of different parts of the human brain, he finished with a rhetorical flourish, declaring that since the different jobs of different parts of the brain were almost worked out, all that remained was to find the part of the brain that co-ordinated and gave meaning to their activity. Exasperated, Maggie raised her hand and asked, mischievously, whether he had perhaps thought of investigating the pineal gland. Ignoring some scattered chuckles, the speaker dutifully noted it down as a possibility, obstinately missing the sardonic reference to Descartes, who believed the pineal gland to link our material bodies with our immortal souls. But we should perhaps sympathise with the neuroscientist and the AI engineer, since fully embracing the "extended mind" is sometimes rather like reaching the end of a sci-fi novel where 200 pages of firm technologic plot dissolves into a brief
befuddling swirl of opaque and fulsome prose, warping space and time as it purports to pierce the mind of God.

Annie Clarkson's 'The Rooms' takes us on a very different journey and, like Millikan's, it is one that meshes meaning with work and with biology. Her protagonist, like many of us, is looking for a job that rewards us with some authentic meaning as well as remuneration. But Hannah’s job is phoney, and the girls that Hannah encounters in the The Rooms are fakes, working jobs that have been carefully and deliberately designed to be meaningful only for their customers.

It may be true that no-one thinks an algorithm will really burn the neighbourhood down, but the notion that following rules and regulations at work can ensure that a real job gets done is uncontroversial: 'A job is a job'. Charles Babbage, the designer of the very first automatic computer, was taken with the idea that a factory could be thought of as a huge machine, with people and computers and lathes and presses all simply its mechanical cogs, getting the job done by going through their motions. But the artisans capable of machining components fine enough for his Analytical Engine design were not minded to play their part in this scheme and the computer went unbuilt. More recently, attempts to incentivise and regulate the caring professions have raised the question of whether formal procedures, metrics, tests and rules can have the 'causal properties' of ensuring that we, our parents, and our children will actually and really be protected, be cared for, and be taught, rather than merely processed. Famously, the oldest profession is also a caring profession, and the tension across the gap between real intimacy and merely 'going through the motions' is surely at its heart.

Like brains, lungs and (according to Millikan) sentences, sex also has its proper functions. Sex is for fun and, ultimately, for reproduction. But these links are deliberately broken inside the Chinese Rooms. Can they be mended? Do Myla, Alina and their co-workers even have reproductive systems behind their sex organs? Do they, for instance, menstruate? Could the 'moon cycle' that Hannah and Samuel discuss, come to have meaning for them? Has it already? The story culminates with Hannah choosing a route forward. Will she succeed? Can we make meaning for ourselves? Or are we subject to the sense and senselessness imposed by the external world. No matter how closely Myla and Alina learn to mimic Hannah's behaviour, they are never expected to actually begin to experience the emotions that drive it. That wasn't part of their program. But could learning, like evolution, ground behaviour and confer meaning on what was initially meaningless? Just as some habitual cycles engender a sense-making autopoiesis, perhaps new kinds of meanings can be co-produced in Chinese Rooms.