ARTICLE

What computers have shown us about the mind

Padraic Monaghan, James Keidel, Mike Burton and Gert Westermann investigate

Over the last half century or so, artificial intelligence models have failed to match the flexibility and adaptability of human performance. However, by incorporating statistical learning and interactivity into modern computational models in the form of neural networks, psychologists are able to gain insight into how our minds operate across a range of cognitive tasks.

This article considers several of these tasks, namely reading, face processing, cognitive development and brain injury, in order to give a snapshot of the range of techniques and questions addressed by researchers using computational models in psychology.

questions

Which aspects of the brain's structure are important for understanding the mind's function?

Are there limits for computers in performing cognitive tasks? Is it a matter of processing 'power', or is the type of processing fundamentally distinct?

resources

www.lancs.ac.uk/staff/monaghan

http://grey.colorado.edu/emergent/index.php/About_Emergent

Churchland, P.S. & Sejnowski, T.J. (1994). *The computational brain.* Cambridge, MA: MIT Press.

With the advent of modern computing in the 1950s, there was an enormous amount of optimism about how quickly and effectively computers would be able to accomplish many of the tasks conducted by humans (Turing, 1950). Skills such as language comprehension or visual object recognition, which are learned early and almost effortlessly by human infants were early targets for constructing competent models, with potentially lucrative outcomes in industrial applications. However, almost 60 years later, we are still waiting for truly effective computer models that can, for instance, translate between languages sensibly, or recognise speech effectively, or identify faces accurately.

These large-scale computer modelling efforts have been driven by the aim of finding a model that works, usually without regard to how humans solve a specific task. But the failure of artificial intelligence models to match human performance provides us, as psychologists, with insight into the way in which the mind is actually solving these tasks. For example, speech recognition systems can become reasonably accurate when they are tuned to a single voice speaking in a regular way within a fairly constrained context. If one or other of these external constraints is not present, then difficulties can arise. So what does this mean for human processing? It means that adaptability and flexibility with regard to contextual information is constantly being utilised by our minds. Clearly, determining how this adaptation to context is accomplished – whether it is the context of the speaker's voice, or the topic of conversation – is of great importance to understanding the cognitive system performing complex tasks such as speech recognition.

The properties of adaptability and flexibility are evident in computational systems that instead of depending on fixed sets of rules, as in early attempts to simulate human behaviour, actually incorporate statistical learning and interactivity in their functioning. Such properties are hallmarks of systems in which information is distributed and interacting, and so the metaphor of the neural network, as implemented in the neural structure of the brain, has been an attractive starting point for many current computational models of psychological processes. Artificial neural networks were originally inspired by the neural architecture of the brain in terms of sets of interconnected neurons transmitting signals via axons, dendrites and synapses between them (McCulloch & Pitts, 1943). In the brain, neurons that are co-active tend to increase the strength of their interconnection (Antonov et al., 2003; Hebb, 1949), and so the statistics of the environment and the task can be incorporated into the neural system itself. Artificial neural networks, then, instantiate the computational principle arising from the brain's cellular structure in terms of reflecting the statistical properties of the task, by employing many small interconnected processing units and adapting the strength of the connections between these units.

In the remainder of this article we provide a set of examples of how this neural network approach to exploring how psychological processes can be implemented in the brain has revealed a great deal about how our minds operate across a whole gamut of cognitive tasks. The next two sections provide two cases in which our understanding of brain function has proceeded in tandem with developing computational models of complex tasks: reading and reading impairment, and then

references

Antonov, I., Antonova, I., Kandel, E.R. & Hawkins, R.D. (2003). Activity-dependent presynaptic facilitation and hebbian LTP are both required and interact during classical conditioning in Aplysia. *Neuron, 9*, 135–147.

Bruce, V., Henderson, Z., Greenwood, K. et al. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*, 339–360.

Burton, A.M., Jenkins, R., Hancock, P.J.B. & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*, 256–284.

Coltheart, M., Rastle, K., Perry, C. et al. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256.

Galaburda, A., Menard, M. & Rosen, G. (1994). Evidence from aberrant auditory anatomy in developmental dyslexia. *Proceedings of the National Academy of Sciences, 91*, 8010–8013.

Hebb, D.O. (1949). *The organization of behavior: A neuropsychological theory.* New York: Wiley.

Jenkins, R. & Burton, A.M. (2008). 100% accuracy in automatic face recognition. *Science, 319*, 435.

Keidel, J.L., Welbourne, S.R. & Lambon Ralph, M.A. (2010). Solving the paradox of the modular and equipotential brain: A neurocomputational model of stroke vs. slow-growing glioma. *Neuropsychologia*, doi:10.1016/j.neuropsychologia.2010.02.019.

McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin*

visual processing (in particular face recognition). Then we discuss the advantage of neural networks in reflecting human development, in terms of the adaptability and flexibility in the way in which these models learn. Finally, we report with more detail a further advantage of this approach to describing and understanding impairments to psychological functioning as a consequence of brain injury.

Reading

There are many different levels of language structure that can be modelled: from discourse, where topic and style are considered, right down to the interpretation of acoustic events as speech sounds. In this section, we focus on just one level – that of learning to read words.

In order to recognise a word, we need to be able to identify the letters, and convert the letters into speech sounds. An early computational account of how this may be achieved assumed that our minds apply a set of rules about which letters make which sounds. So, the letter 'b' is always pronounced /b/, the letter 's' is pronounced /s/, unless it is followed by 'h', in which case it is pronounced /ʃ/, and so on. Such rule-based systems are evidently useful, and can support children's reading development in many cases, as attested to by phonics training. However, some letters are pronounced in irregular and largely unpredictable ways, such as the 'i' in 'pint', which is usually pronounced differently in similar contexts as in 'tint', 'lint', 'mint', and so on. To deal with these exceptions to rules, such models proposed in addition that the reading system also contains word-level representations with a stored pronunciation of the whole word (e.g. Coltheart et al., 2001).

We are still waiting for truly effective computer models

However, an alternative is to consider that the brain is learning not a set of rules for converting letters to sounds, but rather the statistics of the relations between certain letters or sets of letters, and certain sounds or sets of sounds. There is, then, no distinction between reading an exception word compared to reading a regular word, it is just a matter of the degree to which the statistical associations used by the model are regular in forming the letter-sound mapping. Models that have employed neural networks to learn these statistics are more successful in reflecting the graded effects of various levels of regularity in spelling–sound correspondences in words and nonwords (Zevin & Seidenberg, 2006).

Neural networks can couple this learning of statistical regularities with gross anatomical constraints in the brain (e.g. in terms of how visual information is integrated in the reading system) to solve the problem of how we learn to read. Such models have investigated how the left and right hemisphere division has an impact on when and where the visual information must be combined in order to read effectively (Monaghan & Shillcock, 2008). If the two halves of the model do not integrate their visual information at an early enough stage, then dyslexic behaviour emerges in the model, consistent with accounts of hemispheric dissociation in some dyslexics (Galaburda et al., 1994).

The model can also implement other accounts of developmental disorder resulting in dyslexia, for example by simulating accounts of dyslexia in terms of disturbance of visual input, or phonological impairments (see Monaghan & Shillcock, 2008, for a review). The benefits of computational modelling mean that the different natures of these impairments and their behavioural manifestations can be explicity compared.

Face processing

The study of face recognition has always been closely associated with modelling. This is partly for the standard psychological reasons (modelling helps theoretical development) and partly because of the desire to automate face recognition for security and surveillance. Engineers aiming to build working systems have to solve the same problem as human observers: how to associate two pictures of the same person, when those pictures may be superficially very different.

In fact, neither problem is solved. Brain imaging and electrophysiological studies have identified some of the neural systems involved in recognising faces, but how this is achieved remains a mystery. On the engineering side, newspapers continue to announce pilot schemes to implement automatic face recognition to enhance security (e.g. on high streets, banks and airports), but fail to report the results of these schemes because they never live up to the early expectation.

This is an area in which engineering could take more notice than it does of psychological results. At first pass, the problem of face recognition is simply stated: We need to store a set of photos in a database. We then need to take a new photo of someone (for example at a border crossing) and match it to our database. If we have a sufficiently clever matching algorithm, it should be possible to establish whether the person is in the database, and if so who it is. However, it turns out that this is a very difficult task, and one which humans cannot do reliably. In the last 10 years it has become clear from a great many studies, that viewers are surprisingly bad at matching two photos of an unfamiliar person, even when the photos are of very good quality, and taken minutes apart (e.g. Bruce et al., 1999). We are perhaps misled by our own competence here. It turns out that we are extremely good at recognising faces when they are familiar to us, and can do so even

of Mathematical Biophysics, 7, 115–133.
Monaghan, P. & Shillcock, R.C. (2008). Hemispheric dissociation and dyslexia in a computational model of reading. *Brain and Language, 107*, 185–193.
Rogers, T.T., Lambon Ralph, M.A., Garrard, P. et al. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review, 111*, 205–235.
Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.
Zevin, J.D. & Seidenberg, M.S. (2006). Consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language, 54*, 145–160.

in severely degraded images. This expertise perhaps leads us to the false conclusion that we are good at face recognition generally. We are not.

Recently, studies have begun to ask whether it is possible to re-cast the problem of automatic face recognition to be consistent with human capabilities. Instead of concentrating on ever more sophisticated matching algorithms for comparing two individual photos, an alternative is to instead build into a computer something which captures familiarity in human perception (Burton et al., 2005). This approach explores the possibility that automatic recognition may be improved if, instead of saving individual photos of a person, an abstract representation is stored, derived from a statistical analysis of many instances. In fact, computing a very simple average appears to capture this 'learning' very well, and can lead to very substantial improvements in automatic face recognition (Jenkins & Burton, 2008).

Development

In the past decade a new field of developmental cognitive neuroscience has emerged that makes links between brain development and cognitive development in infants and children. Computational modelling has made a vital contribution to this field by providing explanations of how changes in the brain can lead to changes in a child's behaviours and abilities. A large part of developmental psychology is about investigating at what age children have which abilities, and arguably the main challenge is then to explain why and how these abilities change and develop, that is, to identify the mechanisms underlying cognitive change. Connectionist models, which are (loosely) inspired by the functioning of brain neural networks, are ideally suited to exploring these mechanisms. Candidates for aspects of brain development with relevance to behavioural change are changes in how individual neurons process signals, the wiring up of brain networks according to experience, and the integration of different brain systems into interacting networks that affect each other's functioning.

Of particular interest is the development of categorisation abilities in young infants using modelling techniques. Infant categorisation is often studied by presenting pictures on a computer screen and measuring how much time infants spend looking at each picture. The underlying assumption is that infants look longer at novel, unusual stimuli than at familiar ones. In a typical familiarisation study infants are shown a sequence of pictures of objects from one category (e.g. cats) until their looking time decreases. Then they are shown a new object from the familiarised category, such as a new cat, and an object from a different category, such as a dog. In this case researchers have found that even three- to four-month-old infants indeed look longer at the dog, indicating that they have formed a category for cats that excludes dogs. The looking behaviour of infants has been modelled in neural networks by using models that have to learn to re-generate on the output side what they see on the input side. The idea here is that the model, like the baby, builds an internal representation of the observed object, and the more unusual this object is, the longer the looking time, and the longer it takes to train the model to recreate an accurate representation of the input.

Four-month-olds have been shown to base their category formation on the features of objects (e.g. shape of the head, tail or legs), but 10-month-olds are also sensitive to which features occur together (e.g. a specific shape of head with a specific tail). Therefore 10-month-olds found drawings of animals that contained previously seen features, but in novel combinations, surprising, whereas four-month olds did not. Modelling research shows that a change in the way neurons process information could explain this development. Visual neurons in the brain have an associated receptive field, which is the area of the visual space in which a presented stimulus activates the neuron. These receptive fields shrink with age, possibly based on visual experience. By including this change in a neural network model of object categorisation the model provided a precise account of how changes in neural processing can explain behavioural change during the first year of life.

Brain injury

The principles of adaptability and flexibility of the brain's functioning are observed par excellence following brain injury. If you wished to understand how a computer works, taking off the cover and simply watching it run could only tell you so much. Instead, you would be better served by removing individual parts and observing the effects. Unfortunately, though, computers do not react particularly kindly to such treatment. The function of a computer is very often an all-or-nothing proposition: one minute you are typing up a document and the next you are confronted with your chosen operating system's method of telling you how horribly wrong things have gone.

Compare this with the brain's reaction to damage. At the most general level, we are struck by its ability to maintain a significant degree of function in the face of extreme injury. Further, the impairments that result are neither random nor a simple matter of on versus off, as in a computer. Instead, the specific patterns that result from different types of damage tell us a lot about normal brain function. One of the key insights of connectionist modelling has been how systems with distributed representations account for the varying patterns of cognitive impairment observed in different patient groups.

For instance, Rogers et al. (2004) explored the pattern of impairment that results from semantic dementia, a syndrome associated with bilateral anterior temporal lobe atrophy. These patients exhibit a progressive loss of knowledge concerning objects and their properties, and this loss follows characteristic patterns including an initial loss of specific characteristics of objects (e.g. that a camel has a hump), and overextension of general properties of a category to all members (e.g. drawing four legs on a bird). After training and introduction of damage, the model displayed patterns of damage across multiple tasks highly similar to those observed in semantic dementia patients. The insight available from modelling this behaviour was that information about word meanings was stored in a distributed manner such that meanings gradually eroded as damage increased.

Alternatively, localised damage can sometimes result in rather selective impairments; for instance damage to the right fusiform gyrus is associated with prosopagnosia, or 'face-blindness'. The specificity of these deficits, coupled with the apparent inability for other brain regions to fully restore normal function, has led to a modular view of neural architecture. On this view, the brain is composed of a set of domain-specific encapsulated processors that efficiently perform single tasks, a theory that seems to fit quite well with the effects of acute damage such as stroke.

Recent investigations into the consequences of slowly expanding brain damage have also benefited from connecting brain damage to patients' behaviour via insights available from computational modelling. The sequelae of low-grade glioma (LGG), which are slow-growing brain tumours, has greatly expanded our understanding of the potential plasticity in the adult brain, and how these can alter neural structure. Though LGG often cause damage equal to

Haemorrhagic stroke damage. Coloured magnetic resonance angiography (MRA) scan of the brain of a 68-year-old woman two years after she suffered from a ruptured aneurysm (purple). At the most general level, we are struck by the brain's ability to maintain a significant degree of function in the face of extreme injury.

or greater in extent than that observed in stroke, they typically have only a minimal effect on cognitive functioning. This leads to a seeming paradox: if the adult brain is so plastic, as observed in LGG, why is recovery from stroke often so poor?

Recently, a connectionist model designed to account for the effects of acute versus gradual damage to the brain has tackled this question (Keidel et al., 2010). There are three key factors posited to account for the greatly varying cognitive and behavioural outcomes observed in different types of brain lesion:

- the age at which the damage occurs;
- the speed at which the damage progresses; and
- the existing pattern of connectivity in the brain.

To illustrate how these principles interact, Keidel et al. (2010) employed a novel dual-stream architecture, in which two parallel subnetworks had full internal connectivity but only sparse cross-connectivity. This enabled the model to develop modular processing within each subnetwork but also with the facility to distribute processing across the two subnetworks. The age of the model was implemented in terms of gradually reducing its flexibility to subtly adapt its performance. This was achieved by increasing entrenchment in the model's processing, by biasing the model to produce binary (0 or 1) values at the model's output instead of a more graded range of outputs. Each of the subnetworks was trained on a single task, forming mappings between certain inputs and outputs to represent the essence of a range of tasks, such as recognition of words or recognition of faces from visual input. For the model, there was no overlap between the two tasks, so the inputs and outputs were distinct in each case.

Because of the sparse cross-connections, it would be possible for the two subnetworks to share resources in solving the two simultaneous tasks. But this is not what happened. At the end of the initial training phase, before impairing the model, the full model composed of the two subnetworks had mastered both of the training tasks to a level of 100 per cent correct. Interestingly, removal of all of the sparse cross-connections (equivalent to about 25 per cent of the model's representational capacity as measured by number of links between units; note these links were not removed for the simulations described below) had no effect on network performance. Thus, a form of emergent modularity resulted in the model, even though the possibility for interdependence existed in the model's structure.

To see how this might help account for the varying recovery profiles in stroke and LGG, Keidel et al. (2010) introduced two types of damage into the model. For the stroke simulation, they simply removed the resources within one subnetwork that formed the mapping between input and output representations, meaning that the input and output layers could only communicate via the other unlesioned subnetwork, using the sparse cross-connections. After significant retraining, the stroke simulations were only able to recover to a performance level of about 70 per cent correct on the lesioned task, whilst performance on the unlesioned task remained perfect. Thus, as is observed in the patient population, acute damage yielded a specific deficit that could only be partially ameliorated through relearning, as in cases of strokes affecting the right fusiform gyrus resulting in prosopagnosia.

To simulate the effects of LGG, Keidel et al. (2010) introduced a gradual decrement in the resources available to map between inputs and outputs in one of the subnetworks. This manipulation had the effect of slowly reducing the resources to zero, at which point they could have no contribution to processing in the model. Unlike the result of the stroke simulation, the LGG simulation was able to adapt to the gradual damage, and at no time did performance dip below 90 per cent. After extended relearning, it was possible to remove the entire lesioned hidden layer with only a negligible effect on performance. The computational model thus simulated the effect of instantaneous versus gradual impairment to brain tissue, highlighting that LGG is less catastrophic in terms of loss of function due to the interactivity and distributional nature of processing in the brain.

Conclusion

What each of these different examples of computational approaches has in common is the property of interactivity as a fundamental for brain functioning, which is reflected in connectionist models of the brain's processing. Interactions increase the complexity of the system, but also provide great benefits in terms of adapting to novel circumstances, whether those are externally imposed by the environment, or internally generated, in terms of brain damage.

Though we are a long way from achieving the ultimate aim of simulating human performance in all its complexities, we hope that this snapshot has shown how computer models that incorporate aspects of human processing into their structure and representations have provided enormous insight into the way the human mind processes information.

Padraic Monaghan
is Professor of Cognition at Lancaster University
p.monaghan@lancaster.ac.uk

- **James Keidel**
  is a Research Officer at Bangor University
  pssc08@bangor.ac.uk

- **Mike Burton**
  is Professor of Psychology at the University of Glasgow
  mike@psy.gla.ac.uk

- **Gert Westermann**
  is Professor in Psychology at Oxford Brookes University
  gwestermann@brookes.ac.uk