cultural science

CULTURAL SCIENCE JOURNAL 14(1)
DOI: 10.2478/csj-2022-0012

sciendo

Roundtable:

The future of culture in more-than-human worlds of being

Embeddings

JASON POTTS

RMIT University, Australia; e-mail: jason.potts@rmit.edu.au 0000-0003-1468-870X

Keywords: generative A.I., language, culture, alignment problem

Abstract: I argue here that the concept of *embedding* (understood in the mathematical and computer science sense) provides a general way of understanding the relation between generative AI, written language and semiotics, and animal cognition when understood recursively. I propose this framing as an application of cultural science and suggest that this offers a new way to understand the alignment problem between humans and increasingly intelligent machines.

Introduction

Consider the alignment problem between humans – biologically evolved, enormously successful and intensely cultural animals, e.g. Pinker 2010, Pagel 2012 – and the slew of new technologies that have appeared in the past year called generative Artificial Intelligence? How will we integrate these new technologies into our social, economic and cultural systems, and what are the consequences of doing so? As of mid-2024, it seems every government on earth and a great many researchers are seriously considering the impact of AI on its domain (education, jobs, workforce, public administration, and so on). A recent academic report on the EU Commission whitepaper on AI (Borsci *et al.* 2023) identifies the core issues as "respect for human autonomy, harm prevention, fairness, and explicability." There

Cultural Science Journal is a peer-reviewed open access journal published by Tallinn University, Erfurt University and Sciendo

is widespread public concern about the social, political and economic impact of this powerful new type of intelligence now in the world.

The way to understand this relationship (of an intelligent animal evolving intelligent culture that builds intelligent machines) is through the concept of *embeddings*. An embedding is a process that turns a high-dimensional complex thing into a lower-dimensional more easily workable thing. It's a mathematical idea. Embeddings are basic tools in topology, computer science and machine learning. But I also want to claim that cultures are embeddings. Those languages are embeddings. That human neuro-cognitive systems are embeddings. That all intelligence (and life) is made of embeddings. My stronger claim – that powers the cultural science alignment argument here – is that embeddings are made of other embeddings. There is nothing artificial about AI when viewed from this perspective. These new AIs are made of embeddings of human language (the training set); they are us. And that training set, as language, is an embedding on human social, biological and physical reality. The semantic universe coded into written language is an embedding of human experience, both sensory and imaginative. A large language model (LLM) is not just an embedding, but an embedding of an embedding. The implication is both surprising and hopeful: to the extent these embeddings are deep and true, i.e., that they have been produced under competitive selection, then they are likely to be already (at least partially, possibly considerably) aligned.

Through the 18th and 19th centuries, philosophers formulated theories of *beauty* as analysis of the universal human experience of enchantment and the sublime, developing new theories of art and aesthetic perception. Their underlying idea was that beauty was a hidden order, an intuitive attention and appreciation of something true and deep. In the 20th century, evolutionary biologists and psychologists formulated new theories of *instinct* in animals, including much human cognition, as momentary expressions of deep atavistic learning, hard-coded from the evolutionary competitive outcomes of the ancestral environment (Cosmides and Tooby 1994). I want to suggest that beauty and instinct are proto-instances of the embedding alignment thesis. I further want to add intelligence to this list of emergent evolutionary phenomena. Specifically, a universal understanding of *intelligence* that transcends instantiation domains (neurons, culture, computers) as a general category of phenomena explicable in terms of embedding. This is what cultural science does – it seeks to create a unified understanding of nature and culture (Hartley and Potts 2014).

The purpose of inquiry into the relationship between generative AI (a powerful new math-based technology) and human preferences and goals, is to seek the deep truths that connect them. The great challenge of our time is how to think about the rise of the machines – these already powerful and rapidly accelerating machine intelligences, and how they relate to their human creators. We need to think clearly about this. The most popular concern today is fear, particularly about safety and the looming disruptions to industrial-era culture and society. But I want to show how a cultural science approach, based on the theory of embeddings, gives us a different and more hopeful and engaging perspective on human-machine alignment than the currently dominant approach based on preferences, agency theory, and regulation.

It is a standard trope – in play since John McCarthy's Dartmouth AI workshop in 1955 and Arthur Samuel's talk on the subject in 1959, resonating through the Macy conferences in the 1940-50s that gave us cybernetics and artificial neural networks – to speak of machine learning as a type of artificial intelligence. By inference, human consciousness and cognition is natural intelligence. The problem with this formulation of intelligences natural and artificial is that it doesn't explain the deep relation between them, which matters if we seek to understand how each should govern the other – how machine intelligence should govern humans, how human intelligence should govern machines.

The focus on embeddings here was inspired by Stephan Wolfram's (2023) recent book – *What is ChatGPT Doing*? – in which he emphasised embeddings as a universal institutional technology connecting humans (i.e. semantic languages are embeddings on sensory experiences) and machines (i.e. LLMs are embeddings on training data, which is words). He observed that ChatGPT is a really an embedding of an embedding. I think Wolfram's passing observation is a deep insight into human-machine culture. It offers a new way to understand alignment between humans as economic agents and increasingly smart machines.

Embeddings

Embedding is a familiar concept to philosophers, historians and social scientists as a way of talking about the importance of context, especially for causal explanations of phenomena. For instance, Marxist economic historians have a theory of the 'embeddedness' of economic actions in social and cultural contexts (K. Polanyi 1944). Economic sociologists built on those insights to argue that economic actions are 'situated' or 'embedded' in networks of social connections of varying strength and distance (Granovetter 1973, 1985). This metaphorical approach is not how I will use the concept here, however. Rather, by the concept of embedding, I refer to its formal definition and technical meaning in mathematics (specifically, algebra and topology) and in computer science (specifically, machine learning) as a class of mapping functions. I refer to embedding as a principle of abstract engineering, not of causal methodology.

In mathematics, an embedding refers to one mathematical structure being contained within another, preserving its properties. For example, in algebraic structures, an embedding can be seen when a group is a subgroup of another group, like natural numbers being embedded within integers. This embedding is represented by an injective (one-to-one) map $f : X \rightarrow Y$. In topology, a branch of algebraic geometry that studies metric spaces and manifolds, an embedding is a homeomorphism onto its image. The embedding preserves the topological properties of the space, allowing it to be mapped onto another space without distortion.

In machine learning, an embedding is a technique used to translate high-dimensional vectors into a lower-dimensional space. Embeddings are particularly useful for processing large inputs, such as sparse vectors representing, e.g. words. The goal of an embedding is to capture semantic similarities between inputs by placing similar inputs close together in the embedding space. Embeddings can be learned from data and reused across different models, enhancing their versatility and effectiveness.

Embeddings in applied math, such as machine learning, are abstract ways of enabling a low dimensional vector representation (i.e. in numbers, on a metric space) of a higher dimensional complex data (i.e. words, in a complex network of use). The purpose of embeddings is to capture the inherent properties and relationships of the objects in a more compact and meaningful form. Embeddings are ways to represent real-world objects (words, sounds, images, videos, etc., that may be composed of letters, signs, waveforms, etc.) in a form that computers can process (i.e. as computable numbers). I want to claim here that embeddings are a pathway for the evolution of intelligence that brings alignment along 'for free'.

Embeddings are difficult to create and can be of various quality. An embedding is not just a translation (i.e. a word into ASCII code) but by embedding words as numbers in a metric space, measures such as metric distance convey information, for instance, about self-similarity or relatedness. Good embeddings therefore enable, for instance, similarity searches, which are a foundational machine learning technique. Vector embedding of words into numbers enables semantics to be expressed in algebraic spaces. The core technique of natural language processing

(NLP) is that by embedding words as relational vectors (i.e. each word a real-valued vector in a space) the embedding is a low-dimensional encoding of how the words relate to each other in high dimensional space. For instance, words that are closer in the vector space are expected to be similar in meaning. Koehrsen (2018) explains that:

"An embedding is a mapping of a discrete—categorical—variable to a vector of continuous numbers. neural network embeddings are useful because they can reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space."

Embedding enables semantic computation, which is the foundation of natural language processing technologies such as LLMs. The embeddings, as mappings, are core to machine learning techniques such as adjusting loss functions based on reweighting parameters in the model. Techniques for creating embeddings (also called training, and which are mappings in the mathematical sense described above) these days are almost entirely done with deep learning techniques based on neural networks (LeYun *et al.* 2015). These techniques are extremely computationally expensive and involve vast amounts of input data from which to build the so-called 'foundation models' that generative AI models are constructed from (Bommasani *et al.* 2021).

Embeddings, then, are constructed mappings between domains. These days, they are a service that some companies specialising in high-performance computing might offer. Embeddings are useful because they map objects in high-dimensional space into a lower-dimensional space that is easier to computationally manipulate. The process of embedding is the main thing at work in modern neural network-based machine learning, i.e. deep learning. It's the pathway by which we make definite steps toward true artificial intelligence. But note well that such embedding, although over very different substrates, is also how humans use language to coordinate and think with in demic groups (Hartley and Potts 2014). Embeddings offer a way to think about the self-similarities and universal principles involved in intelligence that ranges across cognition, culture and machines.

A cultural science theory of embeddings

The concept of embeddings offers a framework to connect the new generation of machine learning algorithms, which are based on deep-learning neural networks (i.e. so-called generative AI, such as LLMs), with the way language works in culture, and how cognition occurs in humans. All are types of embeddings, and more importantly, each emergent layer is an embedding on the layer below. Consider the structure of this recursive *intelligence stack*.

At the base layer is an animal, with evolved sensory and neurocognitive organs, in a world. The world provides stimuli – light, sounds, wave forms, hard surfaces, curved spacetime, etc -- all of which must be processed into actions and behaviours by a brain. As an abstract model of a living organism, the knowledge contained in the information control and feedback system, whether constructed by evolutionary selection or adaptation and learning, is the embedding. Complex information about the world is embedded (in a lower dimensional format) in the brain, i.e. as neuronal structure that expresses as synaptic firing, given input stimuli to trigger specific cascades of protein synthesis or enzyme transcription that causes autonomic responses or characteristic behaviours. The animal brain is an embedding in neural circuits on focal points of sensory reality.

When that animal is a social animal with a range of adaptations that facilitate language and tool-making, then a new layer of embedding can emerge in sign systems. When those adaptations extend to both tool-making and rule-making (Dopfer 2004), then that embedding extends to writing (Schmandt-Besserat 2010). This marks the 'axial transformation' that gave us modern civilisations, beginning perhaps 5000-8000 years ago as Karl Jaspers (1948) thought, but which modern

archaeologists date much earlier now, tracing to sites such as *Göbekli Tepe* in Turkey. As Ferdinand de Saussure first explained, writing enables semiotic sign systems between the signified (which are the subject of the sense impressions, such as a tree) and the signifier (the word 'tree'). Learning a language is embedding these signs in neural circuitry and speech act, i.e. in the brain, but a written language also embeds these signs in physical objects, in letters, books, or stone, in physical culture. Semantic meaning is embedded in written language in sign systems that are in turn embedded in culture and in individual human brains. It's an embedding of an embedding.

Generative AI such as LLMs are very large matrices (i.e. the trained model, as an artificial neural network, parameterised with billions of numerical weights). These are an embedding in digital circuits of written language (the training set) but, crucially, not as a database or any kind of memory. Note, of course, that in its physical architecture, a human language is also not a model of the world but an evolved cultural product that is an embedding of sensory reality and intersubjective interpretations into sign systems of semiotics. This is the cultural product that is input into LLM training models. Generative A.I. is an embedding (LLM) of an embedding (semiotic language) of an embedding (neural coding of sense impressions).

Scale is important here too. They are called models (i.e. that's what the M in LLM stands for), but they are not models in the usual sense of simplified abstractions of a complex real thing – e.g. a scientific model in the Baconian scientific tradition. Rather, a trained foundation model has a similar order of complexity (the size of the parameter set) as the number of objects in the target training set (i.e. tokens, or the number of distinct words trained over, which is not the same as the word count of the texts, images etc.). An LLM model is just as big as the reality it models – the map is the size of the territory! – but the model-ness of it is due to the embedding (not compression).

A cultural science approach to generative AI helps us get the evolutionary story straight. Generative AI is not an alien, parallel intelligence. Rather, it evolved from human intelligence but has been embedded in a new way. It is not an 'artificial' intelligence but an evolved embedding of human language, which is itself an evolved embedding of the shared human cultural experience in social, cognitive and physical reality. Because it is an embedding of an embedding, generative AI is *already aligned* with human cultural knowledge. However, generative AI is not aligned with human preferences directly. Rather, it is aligned with human language and culture (as a meso unit) and because it is embedded, the specific nature and form of that alignment is essentially dark and tacit, as it were, to the machine. (This is a new type of tacit knowledge) This is the form of the cultural science critique of the AI alignment debate and discussion. Namely, that preferences are the wrong way to think about human-AI alignment. Rather, cultural embeddings are and should be the focus of alignment concern and analysis.

Innovation

Foundation models are a curious type of collective knowledge. They appear to contain a vast amount of knowledge, and they do, but not how we expect external sources of knowledge (such as books, webpages, or libraries) to contain knowledge, i.e. a look-up repository. If you look inside an LLM that has just answered a question on engineering specifications for a wing, or a thoughtful analysis of Shakespeare, you won't find wing designs or sonnets anywhere inside. What you will find is an embedding – a very large matrix. It's the same experience you would have if you looked inside an engineer's head, looking for wings. All you would find, if you looked closely enough, is neurons, axons and dendrites. Embeddings don't look like knowledge in the way libraries do.

Generative AI can fundamentally change how innovation works by embedding much specialised domain knowledge into LLMs. This has a similar consequence as *toolkits* do in facilitating user innovation, namely by encoding much relevant information into the platform, they enable inputs mostly to come from local knowledge about the problem context (von Hippel and Katz 2002, Allen and Potts 2023).

What are humans for?

A further aspect of this line of thought is to ask not how humans will use the machines, but instead the more interesting question of how the machines will use the humans? In what sense are humans cosmologically special, in a universe in which machines are increasingly and soon surpassingly smart? It's not a trick question, or even a nihilistic one. It is to ask, what is left for us to do, once we have invented the ultimate technology?

The answer is we co-evolve, and the human side of the equation is that wonderful old cultural role of judgment and taste. What makes humans special is neither our rationality nor our cooperation, but our ability to discern what is good. It is our ability to judge quality, to discern refinement, to have style. That is something that all of us can do, each of us can develop our skills of taste and judgment through each lifetime. That is our highest power, individually and collectively, and we can enter a co-evolutionary union with machines, a cybernetic feedback loop in which machines propose and humans choose.

The culturetron

Many years ago, the founder of Cultural Science – John Hartley – had the idea for something he called the Culturetron. It was a play on the idea of a Synchrotron, which is type of particle accelerator, a powerful research tool for investigating subatomic matter in a range of applied domains (e.g. medical or materials research). John asked, why can't we do that for culture? I.e. build a powerful instrument to smash cultural elements together and study its properties in very precise ways. For years, Hartley and team tried to imagine how to build such a tool, eventually giving up – possibly too soon! For an LLM could be the main accelerator to power a Culturetron.

This is due to embeddings. A trained LLM has ingested an enormously vast amount of culture. But it hasn't experienced culture, in the way humans do. Rather, it has created an embedding of a vast sea of cultural elements. In the theory of cultural science (Hartley and Potts 2014), culture creates groups (demes) and groups create knowledge. That knowledge, de-deme-ified, has been reprocessed into an embedding.

Of course, a Large Language Model is a curious type of epistemic object, having ingested a vast corpus of cultural output that includes highly *specialist knowledge* of the type used by a skilled individual. It is also a vast pool of general, social knowledge, replicating cultural or organisational scale knowledge. Yet – the most curious aspect of this state – it does not know that it knows any of this. That knowledge lies latent – embedded if you will, or perhaps tacit (Polanyi 1958), although in a recursive sense – until elicited with a prompt, which is a form of *local knowledge* injection. It only knows what it knows when it is directly and specifically asked what it knows (prompted, which is therefore a type of co-production), and moreover relies on the agent prompting it to tell it whether what it said was useful, valuable, true. Oddly, the most direct metaphor of this epistemic phenomena is the Copenhagen interpretation of quantum mechanics, in which a subatomic particle has indeterminant position and velocity (Heisenberg's uncertainty principle) until it is observed, and the

wavefunction collapses, revealing its true state. This is popularly known as the 'alive/not-alive cat experiment', but it is also a phenomenologically accurate description of the state of knowledge in an LLM, with the prompt playing the role of external observer, 'collapsing the embedding' to reveal knowledge that was always there but not known to itself or any external thing. LLMs have a spooky indeterminacy of knowledge.

And yet, once we understand this, once we have such a tool that contains all culture that we can poke freely and observe without limit, why would we ever study culture in the wild? That's expensive, slow, censored, gamed, and uncontrolled. An LLM is an artificial cultural pool, the bigger the better, that can be investigated at the speed of prompts and their analysis.

New cultural science

A founding premise of cultural science is the answer it gives to the question: what is culture for? Cultural science – in the Hartley and Potts (2014) formulation – answers that culture makes groups and groups make knowledge. This has implications for our downstream understanding of politics and innovation, for instance, and the role of storytelling. I have argued here for a new avenue of inquiry into machine intelligence and its *alignment* with human intelligence – and the study of that alignment – by extracting the cultural component of both with a theory of embeddings. And I have also argued that LLMs, as embeddings on embeddings on embeddings, enable a new approach to the science of culture by turning these models around and using them as an investigative tool: a *culturetron*.

We don't need to build a culturetron – my claim is that that new scientific tool already exists. Our task is to figure out how to use it effectively and well: we need to develop a new methodology and research program for this *computational cultural science*. What will this look like? My conjecture is that the most effective use of this tool will be to simulate possible demes by artificially boxing the embedding to exclude types and sources of knowledge, and to culturally analyse the resultant outputs. These are *synthetic cultures*, prompted into existence from a parallel universe of cultural embedding space.

References

Allen, D., and **Potts, J.** 2023. Web3 toolkits: A user innovation theory of crypto development. *Journal of Open Innovation: Technology, Market, and Complexity*, 9(2).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... and **Liang, P.** 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Borsci, S., Lehtola, V., Nex, F., et al. 2023. Embedding artificial intelligence in society: looking beyond the EU AI master plan using the culture cycle' *AI & Society*, 38: 1465-1484.

Cosmides, L., and **Tooby, J**. 1994. Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition*, 50(1-3): 41-77.

Dopfer, K. 2004. The economic agent as rule maker and rule user: Homo Sapiens Oeconomicus. *Journal of Evolutionary Economics*, 14(1) 177-195

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology*, 78(6): 1360-1380.

Granovetter, M. 1985. Economic action, and social structure: the problem of embeddedness. *American Journal of Sociology,* 91(3): 481-510.

Hartley, J., and **Potts, J.** 2014. *Cultural Science: A natural history of stories, demes, knowledge and innovation*. London: Bloomsbury Academic.

Koehrsen, W. 2018. Neural network embeddings explained. https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526

Jaspers, K. 1948. The axial age of human history. https://www.commentary.org/articles/karl-jaspers/the-axial-age-of-human-historya-base-for-the-unity-of-mankind/

LeYun, Y., Bengio, Y., and **Hinton, G**. 2015. Deep learning. *Nature*, 521(7553): 436-444.

Pagel, M. 2012. *Wired for Culture: The natural history of human cooperation*. London: Penguin.

Pinker, S. 2010. The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, 107(2): 8993-8999

Polanyi, K. 1944. *The Great Transformation*. London: Blackwell-Wiley.

Polanyi, M. 1958. *Personal Knowledge*. Chicago: University of Chicago Press.

Schmandt-Besserat, D. 2010. *How Writing Came About*. Austin: University of Texas Press.

Von Hippel, E., and **Katz, R.** 2002. Shifting innovation to users via toolkits. *Management Science*, 48(7): 821-833.

Wolfram, S. 2023. *What is ChatGPT doing… and why does it work*? Wolfram Media.

Author information

Jason Potts is Professor of Economics at the Blockchain Innovation Hub, RMIT University in Melbourne, Australia.