Jun 25th, 9:00 AM

Metaphors for designers working with AI

Dave Murray-Rust
Delft University of Technology

Iohanna Nicenboim
Delft University of Technology

Dan Lockton
Eindhoven University of Technology

Citation

Murray-Rust, D., Nicenboim, I., and Lockton, D. (2022) Metaphors for designers working with AI, in Lockton, D., Lenzi, S., Hekkert, P., Oak, A., Sádaba, J., Lloyd, P. (eds.), *DRS2022: Bilbao*, 25 June - 3 July, Bilbao, Spain. https://doi.org/10.21606/drs.2022.667

DRS 2022
Bilbao 25th June — 3rd July

Metaphors for designers working with AI

Dave Murray-Rust[a,*], Iohanna Nicenboim[a], Dan Lockton[b]

[a]Delft University of Technology, The Netherlands
[b]Eindhoven University of Technology, The Netherlands

*corresponding e-mail: d.s.murray-rust@tudelft.nl

doi.org/10.21606/drs.2022.667

Abstract: In this paper, we explore the use of metaphors for people working with artificial intelligence, in particular those that support designers in thinking about the creation of AI systems. Metaphors both illuminate and hide, simplifying and connecting to existing knowledge, centring particular ideas, marginalising others, and shaping fields of practice. The practices of machine learning and artificial intelligence draw heavily on metaphors, whether black boxes, or the idea of learning and training, but at the edges of the field, as design engages with computational practices, it is not always apparent which terms are used metaphorically, and which associations can be safely drawn on. In this paper, we look at some of the ways metaphors are deployed around machine learning and ask about where they might lead us astray. We then develop some qualities of useful metaphors, and finally explore a small collection of helpful metaphors and practices that illuminate different aspects of machine learning in a way that can support design thinking.

Keywords: metaphors; machine learning; conceptual foundations; computer science provocations

1. Introduction

Metaphors do many different things. Far from being linguistic games or rhetorical flourishes, they are “pervasive in everyday life, not just in language but in thought and action” (Lakoff & Johnson, 1980). Structural metaphors link concepts together – ‘demolishing an argument’, spatial metaphors provide orientation – feeling ‘up’, ontological metaphors make concepts real - ‘inflation is taking its toll’ (ibid). The practices of machine learning and artificial intelligence draw heavily on metaphors: the neural net as an analogy between a collection of linear algebra and human thinking, or the idea of 'learning' as a shorthand for the operation of a backpropagation algorithm over those networks. These can be high level conceptualisations that shape the approach to systems, or low-level metaphors that shape engagements with the technologies. Computer science as a whole is a deeply metaphorical field. Mathematical formal properties are translated both into low level code that seeks to embody the constructions, and into folk descriptions of the systems in terms of familiar physical systems and objects (DeVito, 2021).

Metaphors have long been part of designers' toolkits (Cila, 2013; Jung et al., 2017). They can help with ways to understand situations, an play into the practice of framing and development of worldviews (Pee et al., 2015). They can be generative of new ideas, constraining and opening technical possibilities. Metaphors form part of the imaginaries of technologies, shaping the way that societies construct the space of the possible (Bory & Bory, 2015; Lockton et al., 2019). In the context of AI, metaphors have recently emerged as a way to give personality to conversational agents, shaping end-user perceptions of what their capabilities are and the relations that should be formed (Khadpe et al., 2020). The "X as Y" style of analogic metaphor can be generative of new ideas, especially around emerging technologies, casting robots as sidekicks (Luria, 2018), or bridges, or choirs (Alves-Oliveira et al., 2021). Dove and Fayard's (2020) work on AI as monster is particularly notable here for a deep connection with the technology, through the lens of history and fable ; there are also echoes of this in Byrne et al's (2021) exploration of concepts from the supernatural being applied to 'invisible' technologies such as AI. This leads to the idea of metaphoring as a practice, within the design process (Lockton, 2020, e.g. 2021), which Dudani systematises, finding associations to other domains of knowledge; embodiments of experience; materialisations of the intangible; diversification to capture facets of complex systems and probes for critical questioning (Dudani, 2021)

In this paper, we are interested in metaphors that shape the way that designers can approach the techniques, technologies and artifacts of artificial intelligence. These conceptual metaphors can easily become invisible. Through pragmatics, 'I'm walking on sunshine' is understood to not be literally true (J. R. Searle, 1993). However, if one says 'his words carry little meaning' it can be forgotten that 'carry' is metaphorical, and so the implicit framing of conversation as a contextless exchange of information purely contained in the words goes unnoticed (Reddy, 1993).

Agre, drawing on Foucalt, Derrida and Husserl, encourages thinking of metaphors in technical practice as being part of the way a particular group develops their worldview – not a transparent set of relations between language and physical reality, but a set of moves as "ideas are made into techniques, these techniques are applied to problems, practitioners try to make sense of the resulting patterns of promise and trouble, and revised ideas result" (Agre, 1997, p. 38). As with any practical understanding, there are parts where the theory holds well, and parts where it breaks down – the *centres* of well behaved, theoretically explained phenomena, and the *margins* of the unruly, peripheral and unconsidered (Ibid, p. 43). The tools that are used in discourse reconfigure the world, constraining and enabling what can be said (Barad, 2003) – "It matters what matters we use to think other matters with; it matters what stories we tell to tell other stories with" (Haraway, 2016, p. 12). This can be seen as well in Inayatullah's Causal Layered Analysis (1998) which delves through litany and narrative to find metaphor as the deepest level of social understanding, potent for futuring.

Overall, when **forging** into new **domains**, it helps to be aware of what the **underlying** metaphors being used are, and which **aspects** of their **structure** carry through. Here, we explore some of the unnoticed, pervasive and **foundational** metaphors that direct investigation and **engagement** with technologies.

2. Metaphors in computer science and AI

> A computer might be constructed so that its operation can be narrated using words like "knowing" or "reminding," but those words have a life that goes far beyond - and may even conflict with - the artifacts that are supposed to embody them. (Agre, 1997, p. 60)

Computer science rests on many metaphors (Blackwell, 1996), that have shaped the ways that the systems are conceived of by designers and practitioners. These are often stacked precipitously: the output of analogue transistors is abstracted to a collection of binary values, whose location is then abstracted to become a file, working up through layers of interface toolkits to windows, with buttons and so on. Sometimes, these abstractions are *leaky*, and parts of the lower levels protrude, whether the abrupt stop of a failing hard drive, or the visual patterns of misplaced write to graphics memory. The ISO model of networks takes a seven layer journey of abstraction from the physical world up to the application, creating the space where HTTP – the protocol that underlies most behaviour on the web – can be constructed. There is a constant precarity to these stacks, a consistent need for each layer to allow the layer above to work without context, without reference to the particularities of the cables, bytes, packets, requests, clicks or even gestures of the level below.

There is a fluidity between understandings of computers and of human thought and practice. Talking about computer memory draws parallels to how humans remember, some of which are warranted, and others less so. Calling a selection of memory a file originally captured the idea of bringing pieces of related information together, and conjured imagery of filing cabinets and neatly ordered assemblages of card and bent metal. As far back as 1995, Gaver (1995) argued that many desktop metaphors in interaction design were in fact no longer really metaphors, the digital meaning having supplanted the original or having become an additional, distinct meaning. There is an intimacy between the language and the metaphor, a link between the sequential operations of instructions and the timeless realm of formal mathematics, that shape the emergences of programming practices.

Artificial intelligence has always been a material practice (Agre, 1997, p. 10), that relies on metaphors to construct the links between human thought, mathematical properties and the exigencies of technological possibility. The very term 'artificial intelligence' is a deeply burdened metaphor, which we will not disentangle here except to note that the field might have gone differently had the less sensationalist term 'complex information processing systems' (Newell & Simon, 1956) been adopted. The emergence of AI alongside advances in cognitive science means that there has been a constant permeability around terms covering

both how to understand the behaviour of computational systems and the human brain – including development of metaphor itself (for example Barnden, 2008). Explicit AI metaphors have set out to do particular things. The 'Chinese Room' argument (J. Searle, 1999) used a metaphor that likened any AI system to person translating a foreign language without access to the world is echoed by the 'Stochastic Parrot' (Bender et al., 2021) to critique the gap of meaning between the system and the world in which it exists.

Other metaphors seek to characterise a field: the phrase 'data is the new oil' implies accumulation through dispossession and extractive tendencies in a new frontier – a wilderness ready for claiming (Thatcher et al., 2016). This can be contrasted with ‘data as toxic waste’ – "dangerous, long-lasting and once it has leaked there's no getting it back" (Doctorow, 2008) – as smog, breadcrumbs or markets (Watson, 2015). The metaphors of Software or Data Carpentry cast computational techniques in terms of familiar material practices of cleaning, shaping and assembling. Algorithms can be seen as mediation, as performance as infrastructure (Wagenknecht et al., 2016), even if the concept of algorithm itself is a simplification of a collection of enactments, systems and translations (Kitchin, 2017).

3. Metaphors in machine learning

Here we pick up some common metaphors used about machine learning systems. Some of these are interesting for their invisibility – in general use, they are unnoticed, seen as descriptions of the phenomenon rather than metaphorical descriptions. Some are interesting for their leakiness, where they break down and mislead – where rather than making machine learning more accessible, they make it more difficult for designers to engage, or the different interpretations of the metaphor exacerbate the gap between different fields hindering multidisciplinary collaborations. Some are interesting for the ways that they implicitly define the roles and responsibilities of humans engaging with the systems, attitudes toward failure and the space that people have to imagine and act.

3.1 Training and Learning

One of the most pervasive metaphors within the practice of machine learning is the idea that models learn. While there is no doubt that models change their behaviour in response to new data, the metaphor of learning brings with it the possibility that they learn in a similar way to humans. This makes it easy to forget that impressive surface level performance does not necessarily correspond with other abilities that humans have, such as generalisation and reasoning. This is an example of a metaphor where some of what is carried across does not hold. In this case talking of 'fitting' a model is a cleaner description of e.g. the application of backpropagation to a network in response to a dataset: it carries the idea that the model is being fitted to a particular set of data, rather than learning a general principle. This also highlights the idea that there is something that is being fitted, inviting questioning of what the process is, rather than the somewhat magical property of learning. If one applied a standard sense of ‘learning’ to a conversational agent, it leads to the expectation that this

particular Alexa learns about the particular interactions, the habits and styles of conversation that these users prefer and so on. It does not point to the extended assemblage that learns from aggregate properties of a multitude of interactions across the globe – an incomprehensibly different form of learning.

3.2 Explanation

An example of an almost unnoticed metaphor is the use of 'explanation' to point at the ways in which information on model operation and decisions is collected and presented. This is not to undermine the project of Explainable AI, but to point out a difference between the computational artefacts produced and the social processes by which humans construct explanation (Mittelstadt et al., 2019). This is analogous to the way that Lane et al. (2018) draw on the metaphor of journeys to describe mathematical proofs not as formal symbol manipulation, but rather crafting pathways for others to follow. Similarly, O'Hara (2020) highlights the difference between having knowledge of something and understanding it, as well as the idea that and explanation – outside of formal argumentative settings – is a process, that unfolds through engagement. Emerging techniques in XAI work in this direction, and expand from practitioner oriented debugging to end-user facing processes that support understanding, but the idea that the model produces an explanation of itself remains widespread and draws thinking in the wrong direction.

3.3 Bias

To say a model is biased has a similar surface meaning to saying that person is biased: decisions are being made irrationally, unfairly, in ways that show unethical systematic preferences. However, as a technical word, it has meanings in different communities: in electronics a transistor is biased to create the neutral point that it operates around; the equation for a straight line is given in terms of slope and bias; in tailoring, bias cuts create additional flexibility and conformity in parts of garments.

Within machine learning, bias has several meanings. On a very fundamental level, it can refer to one of the parameters of an individual neuron in a network -- the activation of a cell is the sum of the inputs times the weights, plus the bias, leading to colloquial expressions such as 'loading the biases' to mean configuring a network for a particular task. This is more than simple wordplay -- the biases are the way that the model encodes its fitting of input data. The biases causally connect the output to the input, and set out what outputs are possible, the very things that make the model useful, with output that is not random.

Further to this, Hildebrandt (2019) works through several different forms of bias, from the language bias that covers what kinds of hypotheses a model can fit to, through biases in the data fed in, biases introduced by labelling (feature space) and the ways that the data relate to the world (ground truth).

All of this is manageable – it is not that designers cannot engage with complex, nuanced ideas – but for the pernicious idea that talking about bias raises the possibility of an unbiased model. This is deep conceptual metaphor, that brings in the possibility of a universal model, devoid of context, somehow pure. This gets in the way of the more useful task of understanding the biases of a model in terms of the ones which are useful or something bad, intended or accidental, due to the shape of the data or the shape of the model. It sidesteps the way that algorithms are situated, and the co-shaping between algorithms and their world (Draude et al., 2019). A better metaphor would maintain the idea that biases are always relative to *something*, and that *something* needs articulation.

3.4 Black boxes

The 'black box' is another term with multiple meanings, whose danger lies in the crosstalk between them. Originating in electronic circuit theory, it described originally a component that could be known by its operation rather than materiality – "a resistor is considered as a 2-terminal black box defined by the relation $v=R_i$, rather than as a physical device made of metal or carbon." (Belevitch, 1962)**.** Within cybernetics, it provided an experimental stance, so that when faced with "a mass of functioning machinery that was not to be dismantled for insufficient reason", one might ask "what properties of the Box's contents are discoverable, and what are fundamentally not discoverable?" (Ashby, 1961, p. 87)**.** This sense of experimentation is missing from many contemporary uses of *black box*, which convey a sense of powerlessness, both to understanding mechanism and to making sense of output. Pasquale (2015) is a notable counterexample, exploring how a black box view of society leads asking questions about inputs and outputs to discover hidden connections. In the context of machine learning systems, black box metaphors imply the possibility that if we could just open up the intentionally sealed box, we would understand what has been hidden from us[1], which diverts attention from asking how this box came to be, leading to Ananny and Crawford (2016) suggesting instead looking *across* the systems in order to hold them accountable.

4. Alternative metaphors for designers working with AI

> "Technical communities negotiate ceaselessly with the practical reality of their work, but when their conceptions of that reality are mistaken, these negotiations do not necessarily suffice to set them straight." (Agre, 1997, p. 30)

We have discussed several metaphors that cause trouble for a variety of reasons: bringing in unwarranted associations, suggesting impossible goals, framing problems in ways that are not generative of solutions and so on. What then makes a good metaphor? Here we suggest several ways to examine a metaphor to decide how applicable it is to a given situation, within the context of design.

[1] Glanville (1982) explores the nature of these boxes, suggesting the idea that "Inside every white box, there are two black boxes trying to get out".

Firstly, it is important to know that it is a metaphor. Several of the terms above are metaphors that present as accurate descriptions of the phenomenon: learning is typically not presented as a way to understand the behaviour of a system of equations, but as what is happening. Simply noticing that this is metaphorical creates space for critical engagement, a separation between map and territory—but to what extent should the designer present this separation to the 'user'? Should designers seek to highlight the abstraction, the fiction, and its metaphorical nature, or should it be glossed over, map presented as territory uncritically?

Secondly, getting to grips with the bounds of the metaphor, both technically and socially. As with the leaky abstractions in describing computational components, metaphors tend to hold more strongly in some areas than others. Calling descriptive output an explanation works better within technical communities, where it builds on existing understandings to extend to particular circumstances. Using 'reasoning' to capture the manipulation of logical terms through rules captures some aspects of human reasoning, but by no means all; in particular, the 'common sense' kinds of reason that humans perform remains the preserve of science fictional AI systems. This can also be a trade-off – asking how much the metaphor covers up compared with its explanatory power. Designers may often be doing this without explicitly calling out the bounds of the metaphors in use—and practically this might result in systems where multiple, otherwise incompatible metaphors are chosen for particular elements to be presented to users, each selected with their bounds in mind. We do not generally worry that a 'cloud' would not really work as a metaphor for holding 'files', after all.

Third, what does the metaphor centre and marginalise: just as there are places where it holds more or less strongly, there are features and phenomena that it illuminates more or less powerfully. Casting an algorithm as a 'blank slate' centres a narrative of user responsibility rather then developer accountability, while casting it as an 'autonomous agent' implies a lack of user control (K. Martin, 2019)

Finally, how useful is the metaphor? What does it enable us—designers, users—to do? There is a connection here to Logler et. al's work around metaphor cards (2018), where they discuss the strength of metaphors based on some key distinctions:

- Shallow vs. deep: does the metaphor provide an immersive space that supports exploration?
- Provocative vs. conventional: does the metaphor provoke critical thinking and reflection, or does it encode existing thinking?
- One sided vs. diverse: when looking at a collection of metaphors, do they cover similar ground, or spread out to articulate multiple points of view?

The next sections discuss metaphors of the form "Models as X" – it is worth noting that 'model' is itself a metaphor, that requires some untangling. Here, we are using it in a relatively concrete manner, to point to the artefacts typically created and made use of within AI and machine learning communities: a large language model that can produce sentences in response to inputs; an emotion detection model, that translates images into descriptors of emotional state and so on. The concept of modelling, and its relation to physical practices of

model making could be fruitfully discussed, but for brevity, we will allow a sense of what a model might be to emerge from the metaphors under discussion.

With this in mind, we present several examples of AI metaphors and defend their usefulness.

Figure 1. Collection of people in the style of Unreal Engine by hotpot.ai

4.1 Models as collections of examples

One way into thinking about what a model does is to look at it in terms of the data that it contains – rather than as a magic black box, as a collection of examples, whether of images, sounds, decisions or sentences. This is a mid-level metaphor – it covers a way to think of an algorithmic process, capturing something of its properties, without drilling deeply into the way it functions, or making broad claims about philosophical implications. It can be manifest in many forms, such as the "Soylent green is people" trope, to capture the idea that models are really a mulch of traces of humans. In some cases, this metaphor is literally true: the vectors in a Support Vector Machine are the actual examples that shape the decision boundary, with the 'learning' process selecting the most contentious examples (Veale et al., 2018).

In more complex models, this is more allusive, but captures the idea that that model needs to be able, somewhere along the way, to represent its training data.

This is a good metaphor as it is simple and direct, but it gives a window into what the model is doing. It allows questioning of model limits, as it immediately raises the question of what happens when a new example appears, and highlights the extent to which the data collection and curation processes shape model behaviour. It connects to complex ideas around privacy -- 'if this model is full of people, can we extract the people?' leads straight to the idea of inversion attacks and data leakage (e.g. Ye et al., 2021)

Finally, it is extremely open to illustration. Memo Akten's *Ways of Seeing*[2] shows models reconstructing based on the set of images that they have seen, which leads into thinking about the ability of models to concatenate and compose. Anna Ridler's *Myriad (Tulips)*[3] demonstrate the work and labour practices around the collection of the examples that give rise to the models. Two of Jake Elwes' pieces are relevant: *Zizi*[4] vividly illustrates how choosing different examples produces work that asks different questions, and *Machine Learning Porn*[5] is a clear demonstration of how any AI content filter needs to contain the very things that it attempts to filter out.

What does this metaphor imply for designers? It provides a way to think about AI during the design process that sidesteps the idea of a magic bullet or an unknowable superintelligence, rather in terms of a set of examples or even "use cases". These can be seen with properties and limitations much like many other models that designers are familiar with, from physical prototypes to concepts such as personas or user stories. From a user perspective, *examples* lead back to considering similarities and differences to one's lived experience, supporting a more insightful and critical perspective on the technologies than dealing with exotic framings.

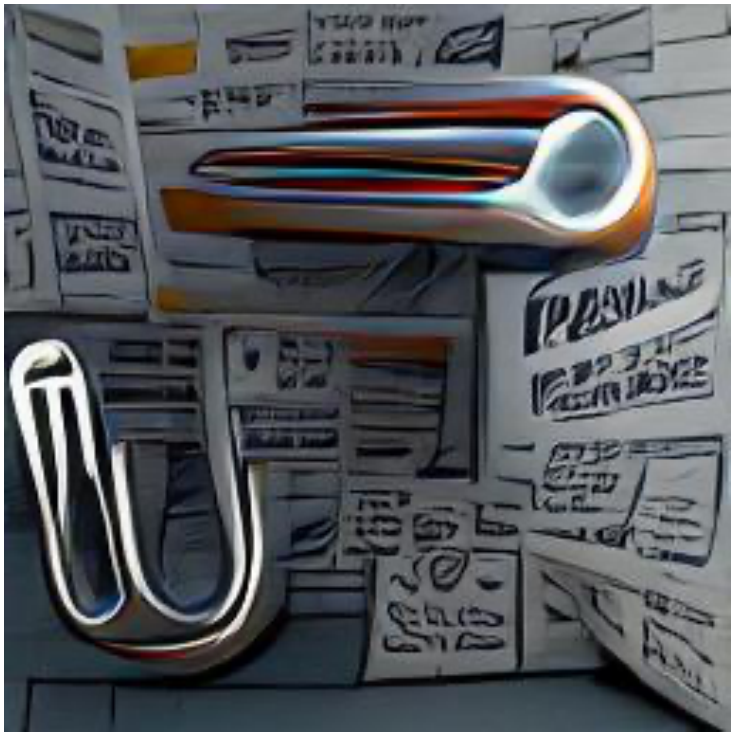

Figure 2. Paperclip maximiser in the style of Unreal Engine by hotpot.ai

[2] https://www.memo.tv/works/learning-to-see
[3] http://annaridler.com/myriad-tulips
[4] https://www.jakeelwes.com/project-zizi-2019.html
[5] https://www.jakeelwes.com/project-MLPorn.html

4.2 Corporations as slow AIs

Another approach to understanding more general uses of AI is to see which existing, familiar things can be described in terms of artificial systems. Stross' metaphor of corporations as a form of slow AI operates on several levels at once(Stross, 2017). The idea that corporations are structurally bound to maximising a particular metric -- shareholder value -- is heavily explored and dissected in European and American culture. This provides a space to bring in Bostrom's paperclip maximisers metaphor (2003) – runaway goal driven systems that seek to maximise their output through increasing intelligence, eventually converting the whole of the universe into paperclips.

Why does this work as a metaphor? Structurally, just as a corporation, AI systems are typically goal-driven, and bring together a collection of components (or employees, or divisions) to handle the work. This provides a point of view to understand the ways that incentives and rewards based on satisfying algorithmic optimisation functions can move a long way from human values (Bridle, 2018). We can draw parallels to the challenge of regulation, as corporate power and functioning has outpaced the ability to regulate its effects, requiring new kinds of legislation, including the question of corporate personhood.

One of the places this metaphor holds well is the sense of distribution and assemblage that can be seen both in the tendrils of even a medium sized corporation and Joler and Crawfords Anatomy of an AI System (2018). It also lets us think into some of mixed systems where humans and AI come together: Wikipedia started as a primarily human driven effort, but is now an assemblage of humans and computational structures, as social processes are gradually formalised and automated, or machine intelligence supports human work in more or less benevolent ways (Tsvetkova et al., 2017).

On a more poetic, yet positive note, Martin (2015) takes up the metaphor of salt marshes to think about what a slow approach to AI might be. Rather than looking at the way corporate structure could be seen as a kind of AI system, it encourages thinking of the entanglements and co-shapings of the various components and actors in terms of a familiar, material ecosystem. As well as a relation to the idea of slow technology (Hallnäs & Redström, 2001; Odom et al., 2021), there is a sense of connection to the traces and practices of past people, the ecology of liminal zones, a connection between the 'social machines' (Shadbolt et al., 2019) of computational coordination and the 'landscape machines' that shape the margins of the sea.

From a design point of view, the “corporation as slow AI” concept has implications for areas such as service design or user experience which mediate the relationships users have with corporations and institutions. Does revealing or making explicit how an organisation learns – and potentially how slow that is – change the way the system is presented to a user? How much distinction is made between employee roles, business processes and algorithmic practices? What would familiar diagrams such as service blueprints or user journeys be like if they incorporated (slow) learning loops—and should that learning be made apparent to the user?

Figure 3. the space inside a model in the style of Unreal Engine

4.3 Models as spaces and terrains

Continuing with geographic metaphors, the notion of latent spaces is one of the metaphors that works both poetically and practically to some degree. Some work needs to be done, in particular to start to deal with concept of spaces, bridging between the mathematical descriptions of hundreds of dimensions, and the practical exercises of iteration and development familiar to designers. Space is a busy word -- it can be thought of as what are all the combinations of different parameter settings one could possibly have, or what are all the ways to arrange three intersecting cuboids, or, after Borges' Library of Babel (Borges, 1941), all the sequences of words in a language. These dimensions are radically different from the spatial dimensions of everyday life – the space of images that can exist in 200 by 200 pixel grayscale images has 40000 dimensions – but ideas like that of distances between things persist. On its own, this is a powerful framing for grappling with complex machinery – it captures the possibilities of things, what could have gone into a model and what did go in. The concept of spaces is a foundation for thinking about how to join models together – what kinds of inputs and outputs do they require, and what needs to be done to relate one to another.

Spaces exist inside models as well. The idea of *latent space* points at sets of variables inside a model that ultimately shape its output (Kingma et al., 2014). As an example, the *autoencoder* family of models attempt to map points in a high dimensional space – perhaps images – back to themselves, but through a pathway that involves a much lower dimensional space. While initially counterintuitive, this allows for a more compact representation of the inputs – the model may be able to reproduce an image close to the original with 20 or 30 numbers rather than 40000. This is technically useful, but it is also conceptually useful. The space cap-

tures something of the essence of the data fed into the model, so talk of *journeys*, *trajectories* and *explorations* start to make sense (see for example the interpolations in Bojanowski et al., 2018). A space with 20 or 30 dimensions also starts to relate to the body, allowing the possibility of physical exploration - see Crawford's (2019) XOROMANCY that maps skeletal position to position within a latent space[6]. The extent of the latent space is a summary of what the model has derived from the inputs, and hence what it is able to represent. The resulting metaphor of a space inside the model helps to ask what are it bounds? What are the landmarks and waypoints within it? How do people produce and represent the space, and what is its affect? How far can one see and what is the journey to distant points like?

How can designers work with this metaphor? Landscape metaphors, including 'fields' and journeys within an imaginary space have been used in design research to explore topics including mental health (Ricketts & Lockton, 2019), and interdisciplinarity (Lockton et al., 2020), but the direct analogues with multiple dimensions in AI models have not so far been well developed from a design perspective. Could spaces be a useful metaphor for designers to think about, design, represent, or visualise AI? At the simplest level, could the common use of informal sketched 2 × 2 matrices and other multidimensional 'spaces' in design processes be an easy way in to thinking about the dimensions of AI for designers? It is, however, harder to think about what this might mean from a user perspective — would interfaces which capture the 'spaces' concept make AI even harder to understand, or could they reveal useful dimensions?

Figure 4. Industrial components in the style of Unreal Engine by hotpot.ai

[6] See video at https://www.graycrawford.com/xoromancy

4.4 Models as Industrial Components

Leaving the poetry, a pragmatic approach to working with AI is to ask "under what situations and circumstances is it OK to use this model?". Here, practitioners have drawn inspiration from the electronics industry with its long history of dealing with powerful but dangerous substances and components, where it is important to represent the possibilities of something; the safe ranges of use, the potential side effects, the strategies for both safe use and optimal use. Gebru et al's 'Datasheets for Datasets' work (Gebru et al., 2018)and Mitchell et. al's related 'Model Cards for Models' (Mitchell et al., 2019) draws very directly on this idea, creating an analogy between the datasets that go into models and either electronic components or industrial compounds, both of which are typically supplied alongside a datasheet that describes their safe usage, precautions to take and particular failure modes.

This is an example of a metaphor that casts the unfamiliar in terms of something still somewhat arcane, but which is a known quantity. Thinking critically about the features of the models and their data is analogous to working with the glazes that make ceramics work both practical and beautiful: they can produce ethereally beautiful, highly robust results, but are dangerous to breathe in, behave unpredictably until well understood and need to be baked at just the right temperature.

Should design education be teaching AI as analogous to electronic components that can be added off-the-shelf to a project? Does Mitchell et. al's approach help designers to understand AI at an appropriate level—a well-informed but nevertheless still 'black box' approach? As discussed above in relation to the bounds of different metaphors, perhaps the 'known quantity' idea positions AI models as something like textbooks, with known limitations: out of date, good coverage of some areas but poor coverage of others, author is known to have had certain prejudices and so on. Then, models can be employed critically for what they offer, combined and managed to ameliorate their shortcomings.

Figure 4. Geofoam and fossils in a landscape in the style of Unreal Engine by hotpot.ai

4.5 Models as Fossil Data

To come full circle back to the extractive metaphors of "data as the new oil", one of the useful metaphors is the sense that language models are typically built on collections of 'fossilized' language – the strata of internet mediated communication, shot through with seams of journalism and literature[7]. This metaphor helps to understand some of the gap between the way that large language models produce output and the way that humans understand it -- the difference between words in flight as conversational moves tied to situations and context, and the production of mechanically appropriate sequences of text. For designers, the idea of AI as a data fossil might imply presenting 'the model' as a fixed, fossilised thing, encoding old prejudices, that is consulted (similarly to an archive) rather than presenting it as some kind of active learning superintelligence. The idea of fossil data helps to capture the contextless nature of most models, distinct from the situated use of e.g. language in active conversation.

4.6 Models as GeoFoam

As one final metaphor[8], the idea of fossil data can be extended to the idea of language models as geofoam. Geofoam is typically expanded polystyrene – oil removed from the ground and reshaped – that is used to provide foundations for bridges and pavements or otherwise fill voids in the landscape. Despite having to explain what geofoam is, it results in a useful metaphor for current AI generated text: it looks like landscape, but it is really a somewhat

[7] This metaphor comes from @morungos on Twitter, with the fossilised language being "traces of our grasp of the world, but [...] not the grasp or the world" - https://twitter.com/morungos/status/1455527973159948294

[8] From @jjvincent on Twitter: https://twitter.com/jjvincent/status/1455194365169700864

uncanny filler, with little weight, that does not connect ecologically to the soil around it. There is something in the metaphor that will be familiar to anyone who tries to converse with a voice assistant about the difference between surface competence at speaking and having a meaningful, well-founded conversation.

AI as geofoam offers its own intriguing challenges for the designer: should the uncanniness be communicated to the user? Does it make sense to preserve or even highlight the weirdness—or does that forever keep consumer-level AI at the level of amusing or irritating gimmick? Does it suggest new kinds of practices around joining and filling, mediating between interactions, finding a voice that is specific and inhuman?

5. Conclusion

This paper has set out some ideas around the use of metaphors when carrying out design with and around AI technology. It maps out some of the ways that metaphors are used generally, and picks out the use of abstractions and conceptual metaphor that characterise discussion around computer science. From this, there is a discussion of the failure modes of a small collection of current terms – training, explanation, bias and black boxes – in terms of misleading associations, unwarranted implications and lack of positive utility. In response, we offer a set of metaphors that shed light on different aspects of AI behaviour at a variety of levels, whether casting models as curated collections of examples, drawing parallels between AIs and corporations, or pointing to the uncanny 'geofoam' quality of text from language models.

The aim of this work is not to provide a single final set of metaphors, but to provide some ways to notice and question the use of metaphor and evaluate the tools in our conceptual toolboxes. It is not an exhaustive set, and should not replace all the existing metaphors, but we offer this set of new metaphors to the DRS community partly as a provocation to open up the space of possible ways to approach the field of AI. We hope to support designers in thinking about the systems that they are building, how design education and research can make use of new framings in projects, including in speculative ways, and in ways which connect to wider questions in philosophy, science and technology studies, and so on.

We have chosen examples which go beyond solely surface-level treatments of how to *present* these invisible systems to users; instead in each case, the metaphors tell us something deeper about what AI is, and how it behaves—its techniques, technologies, and artefacts. Nevertheless, the user-facing aspect is ever-present, and some familiar issues from the earlier days of interaction design become again apparent in this new field—how much of the systems 'behind the scenes' should be presented to users? Does designers' understanding of the technology match the technologists'—and to what extent does that matter? What are the roles of designers and users here, and does AI give greater responsibilities in terms of the need for awareness and critical evaluation? What agency do we have in this field?

Acknowledgements: The project DCODE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 955990. This research is also partially funded by a Microsoft Research PhD Fellowship 2019 granted to Delft University of Technology.

6. References

Agre, P. (1997). *Computation and human experience*. https://doi.org/10.1017/cbo9780511571169

Alves-Oliveira, P., Lupetti, M. L., Luria, M., Löffler, D., Gamboa, M., Albaugh, L., Kamino, W., K. Ostrowski, A., Puljiz, D., Reynolds-Cuéllar, P., Scheunemann, M., Suguitan, M., & Lockton, D. (2021). Collection of Metaphors for Human-Robot Interaction. *Designing Interactive Systems Conference 2021*, 1366–1379. https://doi.org/10.1145/3461778.3462060

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability: *New Media & Society*. https://doi.org/10.1177/1461444816676645

Ashby, W. R. (1961). *An introduction to cybernetics*. Chapman & Hall Ltd.

Barad, K. (2003). Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society*, *28*(3), 801–831. https://doi.org/10.1086/345321

Barnden, J. A. (2008). Metaphor and artificial intelligence: Why they matter to each other. *The Cambridge Handbook of Metaphor and Thought*, 311–338.

Belevitch, V. (1962). Summary of the History of Circuit Theory. *Proceedings of the IRE*, *50*(5), 848–855. https://doi.org/10.1109/JRPROC.1962.288301

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Blackwell, A. F. (1996). Metaphor or Analogy: How should we see programming abstractions. *PPIG*, 8.

Bojanowski, P., Joulin, A., Lopez-Pas, D., & Szlam, A. (2018). Optimizing the Latent Space of Generative Networks. *Proceedings of the 35th International Conference on Machine Learning*, 600–609. https://proceedings.mlr.press/v80/bojanowski18a.html

Borges, J. L. (1941). *El jardín de senderos que se bifurcan*. Sur Buenos Aires.

Bory, S., & Bory, P. (2015). New Imaginaries of the Artificial Intelligence. *Im@ Go. A Journal of the Social Imaginary*, *6*, 66–85.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277–284.

Bridle, J. (2018, June 21). Something is wrong on the internet. *Medium*. https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

Byrne, D., & Lockton, D. (2021). *Spooky technology: A reflection on the invisible and otherworldly qualities in everyday technologies*. Imaginaries Lab ; School of Design, School of Architecture, and Frank-Ratchye STUDIO for Creative Inquiry, Carnegie Mellon University.

Cila, N. (2013). *Metaphors we design by: The use of metaphors in product design*.

Crawford, G. (2019). *Developing Embodied Familiarity with Hyperphysical Phenomena* [Thesis, Carnegie Mellon University]. https://doi.org/10.1184/R1/8427779.v1

DeVito, M. A. (2021). Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 339:1-339:38. https://doi.org/10.1145/3476080

Doctorow, C. (2008, January 15). Personal data is as hot as nuclear waste. *The Guardian*. https://www.theguardian.com/technology/2008/jan/15/data.security

Dove, G., & Fayard, A.-L. (2020). Monsters, Metaphors, and Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3313831.3376275

Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2019). Situated algorithms: A sociotechnical systemic approach to bias. *Online Information Review*, *44*(2), 325–342. https://doi.org/10.1108/OIR-10-2018-0332

Dudani, P. (2021, September 3). Making Metaphors Matter within Systems Oriented Design. *RSD Symposium*. RSD10. https://rsdsymposium.org/making-metaphors-matter-within-sod/

Gaver, W. W. (1995). Oh what a tangled web we weave: Metaphor and mapping in graphical interfaces. *Conference Companion on Human Factors in Computing Systems*, 270–271. https://doi.org/10.1145/223355.223669

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *ArXiv Preprint ArXiv:1803.09010*.

Glanville, R. (1982). Inside every white box there are two black boxes trying to get out. *Behavioral Science*, *27*(1), 1–11. https://doi.org/10.1002/bs.3830270102

Hallnäs, L., & Redström, J. (2001). Slow Technology – Designing for Reflection. *Personal and Ubiquitous Computing*, *5*(3), 201–212. https://doi.org/10.1007/PL00000019

Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.

Hildebrandt, M. (2019). *The Issue of Bias. The Framing Powers of Machine Learning* (SSRN Scholarly Paper ID 3497597). Social Science Research Network. https://doi.org/10.2139/ssrn.3497597

Inayatullah, S. (1998). Causal Layered Analysis: Poststructuralism as method. *Futures*, *30*(8), 815–829.

Joler, V., & Crawford, K. (2018). *Anatomy of an AI System*. Anatomy of an AI System. http://www.anatomyof.ai

Jung, H., Wiltse, H., Wiberg, M., & Stolterman, E. (2017). Metaphors, materialities, and affordances: Hybrid morphologies in the design of interactive artifacts. *Design Studies*, *53*, 24–46. https://doi.org/10.1016/j.destud.2017.06.004

Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–26.

Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 3581–3589.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lane, L., Martin, U., Murray-Rust, D., Pease, A., & Tanswell, F. (2018). Journeys in mathematical landscapes: Genius or craft? In *Proof technology in mathematics research and teaching*. Springer.

Lockton, D. (2020, June 26). *Survival of Things That Fit: Adaptors as Metaphors for IoT Devices*. Designing for the End of Life of IoT Objects workshop at DIS2020.

Lockton, D. (2021, September 3). Metaphors and Systems. *RSD Symposium*. RSD10. https://rsdsymposium.org/metaphors-and-systems/

Lockton, D., Brawley, L., Ulloa, M., Prindible, M., Forlano, L., Rygh, K., Fass, J., Herzog, K., & Nissen, B. (2020, March 1). *Tangible Thinking: Materialising how we imagine and understand systems, experiences, and relationships*.

Lockton, D., Chou, M., Krishnaprasad, A., Dixit, D., La Vattiata, S., Shon, J., Geiger, M., & Zea-Wolfson, T. (2019, December 10). *Metaphors and imaginaries in design research for change*. Design Research for Change Symposium.

Logler, N., Yoo, D., & Friedman, B. (2018). Metaphor cards: A how-to-guide for making and using a generative metaphorical design toolkit. *Proceedings of the 2018 Designing Interactive Systems Conference*, 1373–1386.

Luria, M. (2018). Designing Robot Personality Based on Fictional Sidekick Characters. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 307–308. https://doi.org/10.1145/3173386.3176912

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

Martin, U. (2015). *Thinking saltmarshes*. Harper Perennial. https://ora.ox.ac.uk/objects/uuid:a9e119d0-9286-4798-9046-03e0f9dbc0e4

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Newell, A., & Simon, H. (1956). The logic theory machine–A complex information processing system. *IRE Transactions on Information Theory*, *2*(3), 61–79. https://doi.org/10.1109/TIT.1956.1056797

Odom, W., Stolterman, E., & Chen, A. Y. S. (2021). Extending a Theory of Slow Technology for Design through Artifact Analysis. *Human–Computer Interaction*, *0*(0), 1–30. https://doi.org/10.1080/07370024.2021.1913416

O'Hara, K. (2020). Explainable AI and the philosophy and practice of explanation. *Computer Law & Security Review*, *39*, 105474. https://doi.org/10.1016/j.clsr.2020.105474

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pee, S. H., Dorst, K., & van der Bijl-Brouwer, M. (2015). Understanding problem framing through research into metaphors. *IASDR 2015 Conference*.

Reddy, M. J. (1993). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed., pp. 164–201). Cambridge University Press. https://doi.org/10.1017/CBO9781139173865.012

Ricketts, D., & Lockton, D. (2019). Mental landscapes: Externalizing mental models through metaphors. *Interactions*, *26*(2), 86–90. https://doi.org/10.1145/3301653

Searle, J. (1999). *The Chinese Room*.

Searle, J. R. (1993). Metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed., pp. 83–111). Cambridge University Press. https://doi.org/10.1017/CBO9781139173865.008

Shadbolt, N., O'Hara, K., De Roure, D., & Hall, W. (2019). *The theory and practice of social machines*. Springer.

Stross, C. (2017, December 27). *Dude, you broke the future!* 34th Chaos Communication Congress. http://www.antipope.org/charlie/blog-static/2018/01/dude-you-broke-the-future.html

Thatcher, J., O'Sullivan, D., & Mahmoudi, D. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space*, *34*(6), 990–1006. https://doi.org/10.1177/0263775816633195

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of Wikipedia. *PLOS ONE*, *12*(2), e0171774. https://doi.org/10.1371/journal.pone.0171774

Veale, M., Binns, R., & Edwards, L. (2018). Algorithms that remember: Model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180083. https://doi.org/10.1098/rsta.2018.0083

Wagenknecht, S., Lee, M., Lustig, C., O'Neill, J., & Zade, H. (2016). Algorithms at work: Empirical diversity, analytic vocabularies, design implications. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 536–543.

Watson, S. M. (2015). Data is the New "_". *Dis Magazine*.

Ye, J., Maddi, A., Murakonda, S. K., & Shokri, R. (2021). Enhanced Membership Inference Attacks against Machine Learning Models. *ArXiv:2111.09679 [Cs, Stat]*. http://arxiv.org/abs/2111.09679

About the Authors:

Dave Murray-Rust is an associate professor in TU Delft's Department of Human Centred Design, researching human-algorithm interactions, and developing methods for working with AI futures.

Iohanna Nicenboim is a Microsoft Research PhD fellow at TU Delft, investigating human-AI interactions through more-than-human design. For the past ten years, she has been practicing speculative design to create fictions that highligh the ethics of living with autonomous technologies in everyday life.

Dan Lockton is currently an assistant professor in TU Eindhoven's Department of Industrial Design, with climate futures and design research methods as his focus. He also runs the Imaginaries Lab, an independent research-through-design studio, based in Amsterdam.