Computer Supported Cooperative Work (CSCW)
https://doi.org/10.1007/s10606-023-09483-6

EDITORIAL

Machine Learning and the Work of the User

Richard Harper*[1] & Dave Randall[2]
*[1]*School of Communication and Computing, Lancaster University, Lancaster LA1 4YD, England (E-mail: r.harper@lancaster.ac.uk);* [2]*Siegen University, Siegen, Germany*

Abstract. This paper introduces the collection of the Journal on Machine Learning (ML) and the user. It provides a brief history of ML from the 1950's through to the current time, sketching the nature of the kinds of precursor AI techniques used in such things as expert systems right the way through to the emergence of ML and its tool sets, including deep learning. It concludes with the 'generative AI' used in such ML technologies as PaLM and GPT-3. The history highlights key changes and developments in ML, the especial importance and limitations of deep learning, and the changing attitudes and expectations of users in an environment when ML can and often is oversold. The paper then explores the ways CSCW research has addressed the social context of organisational systems and how the same can apply for ML tools and techniques. It urges research that focuses on the particular ways that ML comes to fit into 'real world' collaborative work sites and hence speaks to the CSCW cannon.

1 Introduction

The history of machine learning (ML) has been a history of two sides: exaggerated hype and unnecessary fears on one, and, on the other, steady, if slow progress in technology on the other. Any attempt to grapple with ML needs to separate these two concerns before it outlines a third concern, the kind of relationship CSCW might have with them. It is our view that a new perspective, relevant to the issues at hand, needs developing. Such a view will, we argue, owe something to CSCW because CSCW is founded on the interactional issues that are sometimes overlooked in the ML literature. However, (and we remain agnostic for the moment), it may be that the peculiar features of ML might require some 'special' treatment.

As we approach these and other possibilities, it is perhaps worth sketching what the issues at hand might be. First of all, there is the technology. This has, as we say, made steady but slow progress over the years, but is, in many ways, relatively straight forward. The basic concept of machine learning was devised in the 1950s; the notions behind deep learning a decade later. Computer hardware that would make these computational techniques powerful was theoretically conceived in the 1980s but only became practical in about 2010.

Published online: 02 February 2024

Timelines aside, certain aspects of this development need appreciating. For example, ML depend in very large part on innovations in data. Over the past twenty year or so, acts by users such as keyboard entries, say, or visual signals from a camera, or webs of datum to do with crowd behaviours on the internet, have been transformed by engineers who sought out the 'ML-relevant features' of the phenomena in question. Extracting these features took years in some instances, but once done has almost become taken for granted, as Crawford (2021). It is also artful, and makes successful ML dependent in human ingenuity (Duboue 2020). Feature extraction (or feature engineering, an alternative but perhaps better term for the way data had to be developed), turns around how data might map its own internal relationships – identifying similarities between instances of data. It is through discovering what the relationships might be and hence their potential for interpretation and 'learning' that determines for engineers the properties that need storing. Acts by individual users, as a case in point, have to be defined as instances of types (or categories) that apply for all users (or at least large numbers of users) and not expressions of individuated acts, and it is these generic types, a selected 'ontology', that allows connections between those acts to be made. The work of feature engineers is to identify these types (and indeed continue to do so as new sources of data capture emerge).

It is perhaps worth noting as well that these types (which are indexes to the sought for patterns) are essentially statements of virtual distance (temporal and spatial) as well as calculated notions of likelihood (i.e., the greater the likelihood, the 'nearer' the interpretation is to the 'right answer', whatever the right answer might be). The nature of this distance keeps evolving with transfer learning, a quite recent development, treating distances as 'learnt weights' and taking these weights, or ''distances as we suggest, from one domain and applying them in another to create a new model of whatever is the phenomena in question.

These measures, whatever they might be, are embedded in the essential data store of ML systems: the tensor. This is a geometric representation of the data in question, describing the relationship between 'objects' or properties (e.g. weighted colours for images, or scaled distances for shapes in those images). It is thus innovation in data and its ontologies that has allowed the possibilities of AI and ML.

Meanwhile, building on these innovations, we arrive at AI and ML applications (or engines) themselves. These consist of various ways of processing data, ways which document patterns such that the relationship between that data and new data leads to the identification of new instances of the pattern. Crucial here is that these newly identified instances can generate further instances or patterns; they can, as it were, predict. These applications are best understood as, above all, innovations in respect of scale and speed (though the terms used to name these applications and their components can evoke startling analogies and metaphors. This feeds the ideology around ML (See Shneiderman 2022). The difficulties that can derive from language terms used in ML go back a long way (see Macdermott 1976).

Crucial here are the data volumes available through the world wide web. It is these volumes that have driven the development of, for instance, deep learning, which can entail the interrogation of vast, dynamically reconstituted data stores. Scale also has its problems, needless to say, such as when the identifying of patterns becomes stuck in a maze of different possibilities, and so techniques have to be devised that judge between patterns. Another, perhaps greater problem is computational burdens that ML techniques generate. A deep learning algorithm can quickly use up all memory on a computer, slow overall processing time and create strains for an operating system which will most probably not been designed with these burdens in mind. The appeal of techniques like transfer learning is less in what they produce than in the economic approach they take to processing. With transfer learning, the prohibitive problem of over straining computers through combining deep learning engines can be partly bypassed. Transfer learning does this through an elegant way of selecting data. And this in some ways returns us to the importance of feature engineering—ensuring that tensors are apposite for the task at hand.

Of course, this is just to foreshadow what is a quite complex history of a diverse set of technological developments and data transformations, but the point is that one needs to know something about these matters in any attempt to explore the view from CSCW. If ML is so dependent on types of data, for example, what does the CSCW view allow one to appreciate about 'data in action'? Does the view tell us anything important about where data comes from or how is it understood at the point of entry? What about data outputs, at the point of use after ML processing has taken place? As the reader will know, the CSCW view *does* say a great deal about these matters and hence ought to be useful to understanding ML (See also Duboue 2020: 6–15). We shall say something about this later; the papers too will speak to these matters.

The point is that there is an ideology around ML, quite separate from matters to do with the technology itself, and what it can and cannot do and which has to do with 'truth' and 'objectivity'. But in making these arguments, it should be clear that CSCW can offer a distinctive (if not the only) lens or perspective on ML. This, we suggest, will involve studies of the work that the technologies do, and the work that people do with them. We shall conclude our paper by returning to this. Indeed, the point of this paper, and hence the collection, is to address and explore precisely these issues. The papers selected and presented here examine both sides of the issues at hand – the nature of how ML functions (and hence questions about the importance of data and such), and, on the other, the expectations of users that are bound up with the hype and the fear. What the papers report is that these matters of technology and expectation play out in particular ways depending on the context. ML is not all the same, with different 'engines' sucking in different data and accordingly outputting different 'insights'; the organisational sites in which these variants of technology land are different too.

And beyond this, what users expect might vary as well. These are all, in different ways, contextual. Before we get to those papers, however, we turn now to look in more depth at what ML is, its history, and as part of that, how the views that come to surround it have taken the form they have.

2 The Working Of Machines That 'Learn'

Concerns over our relationship with 'intelligent' machinery are not new. On the contrary, questions of expectation and how users react to technology given those expectations go back to at the least as far as the earliest incarnations of Artificial Intelligence (AI). Some of the early 'experiments' into the way in which people understood computer programs, for instance, threw up results that surprised many of those who devised those technologies. Although old news by now, they do indicate something important about how the aggregated data associated with ML might be understood. Take Weizenbaum's Eliza, from the early 1960s. When Eliza was 'tested' by a secretarial user, it was treated as if it was providing sensible, 'intelligent' answers (see, e.g., Natale 2019; Bassett 2019, 2021). Another example is the Parry system, a chatbot which consisted of a crude simulation of paranoid schizophrenia. In the case of Parry, professional psychiatrists, were similarly unable to distinguish between it and a human agent at a level better than guesswork. Today, we suggest, users might approach 'natural language' using systems similarly predisposed to find 'truth'. What one learns is that we can be tempted to imply something (possibly several things) about the technologies here. This is a consequence of how we ordinarily make sense of situations that we 'read' as conversational or narrative and is today consequential for the way we understand the outputs of, for instance, Chat GPT3. 'Generative' AI might seem new, but the reactions it creates are not.

3 Framing Claims About Machines That Learn

The hype that surrounds ML can have deep consequences, then, in terms of user expectations. It can also have implications for those who have an interest in developing the technology – making it easier to justify hyperbole. Much of this has come, not very surprisingly, from those with a vested interest in making large claims – the companies whose business depends on their technologies seeming to be transforming: DeepMind in London, for example, and their parent company, Alphabet in Mountain View, near San Francisco. For the most part, the claims do not emanate from the engineers working in the area for the simple reason that the many of these engineers know better than most the practical limits of the technologies in question. Not all, of course. Geoff Hinton, whose team helped devise a way of implementing deep learning on graphics chips is notorious for the claims he makes, ones that would have even DeepMind blush (for discussion

see Marcus and Davis 2019: 43–44). Most often, exaggeration is the hallmark of the management guru, the knowledge management 'expert', and so on. In a similar vein, the fear that the machines are 'taking over', that their 'intelligence' will soon exceed our own, is a product of a different kind of guru, one who knows and understands very little of the philosophy of mind, or the concept of 'embodiment' and how it relates to practical expertise but whose views seem to imply that AI is an almost magical invention. Kurtzweil (2013) is an obvious example here, but one-time director of AI research at Google, Peter Norvig has stated that AI is the most important technology ever invented; under the rubric of AI he also includes ML (see Russell and Norvig 2017).

Returning to the technology, the fact remains that we have not seen the progress that proponents expected. Hinton suggested that all radiologists would be out of a job by the turn of the end of the last decade, for example. The fact that this has not happened says less about inertia in the named profession as it does about the limits of deep learning. In reality, so-called '5th generation' AI (which incorporates ML) has been very slow to arrive. Minsky's observation in the 1960s that 'in twenty years' machines will be more intelligent than us' has proven very wide of the mark. Through the 60's and 70's progress was made in the use of expert systems and, above all, in the use of expert systems which support very special and limited cases of decision-making by human beings but not much more. This success was—and continues to be—substantially the case with GOFAI (good old-fashioned AI), a rule-based approach to AI, that was predicated on the belief that the rules which govern human behaviour can be made visible and formal. Given this, then, machines can (in theory) perform at least as well as people in the tasks in question (for a critique of 'encodingism', see Bickhard and Terveen 1996). This is only the case with narrow, well-defined tasks, however. One might suggest that in hindsight this success has been down to a kind of feature extraction that had already been done in the contexts in question. What we mean is that in certain tightly controlled decision-making contexts, definitions of right courses of action had been tested and made by people in those contexts. (They still are, wherever expert systems are being currently deployed, which still happens even in the age of ML). To put this another way, we are saying that the 'data' and its relevant features were defined so that they were formal, logical, and hence representable.

The difference between contexts where this definitional work has been done and those where human behaviours were (and are) much broader and varied and the context of decision making equally so, is a highly significant one. Apart from anything else, it is in these contexts that ML is meant to flourish. Besides, it is in these contexts that 'culture' gets implicated in important ways. As we show below, behavioural outcomes are fundamentally determined by cultural matters in all their variation and contextual specificity. It is not always possible to reduce these affairs to logic. In any case, ML applications do not learn in ways suitable for this cultural

heterogeneity, nor, in the case do deep learning ones, do they have any internal symbolic logic to refer to. All they have is pattern recognition engines and the patterns in question have to be analogous to the originating patterns in examples used to train these engines. This has enormous potential when it comes to prediction, as it allows a pattern to be, as it were, foreseen in some given data (or 'inferred', the term preferred by ML engineers), but it also means that deep learning tools are constrained. Their capacity to see different patterns is low, if you like, one might even say shallow. We shall say more about this.

What is sure is that though there may be limits to their powers, ML techniques have been very successful in particular domains. This success derives from innovative combinations of new tools and techniques, including deep learning, and in some instances with parts of the symbolic logic approaches of the old AI. Key to this has been undertaking these developments on the basis of constraining the setting of use (the field, or domain), as well as limiting outcomes. By reference to particular case scenarios – the identification of a visual object say, or the meaning of a phrase in a body of text—ML researchers have tested different combinations of tools and came to determine which generate the closest approximation to 'right outcomes'. And as they have done so, so they have also innovated in respect of the data used for these architectures to enable those outcomes. The 'right' interpretation of a spoken or typed phrase in conversation like human computer interaction is measured in terms of plausibility. It is a best fit, not a perfect or right answer. So, the data that goes into a language conversation model includes phrases and their relations to words through syntax and grammar all demarcated by 'best-fit-reasonable-response-in-conversation' criteria.

In doing so, ML has avoided the pitfalls of old AI where it became impossible to define 'right' answers beyond extremely confined scenarios. A way forward for ML has been seeking practical outcomes through using 'best fit' as target outcomes and seeing what happens – i.e., by seeing how acceptable best fit outputs are. It turns out these outcomes are often good enough to be useful. Even better, they can themselves become a resource for refining outputs. Best fit can get better and better. This doesn't mean that ML (and deep learning in particular) cannot make shocking mistakes, but it does mean that by in large outputs are good enough for many tasks. In short, the move from epistemic to pragmatic tests has been key to success and distinguishes GOFAI from ML and its sub technologies like deep learning.

4 The Machineries of Learning

This naturally leads us to what the machinery in ML is. It is essentially nothing more than a set of techniques for pattern recognition. The trick, if trick it is, is that once a set of patterns have been documented inside the application, that same application can then find instances of those patterns in new examples of data. These examples might be isomorphic with the one first used in the training

sets, or 'like' or 'near to' that first example. The system has learnt what 'like' is. Like is a synonym for distance and-or nearness in instances of data.

An important point to recognise, however, is that no ML application learns without human intervention, even though there are two broad types of ML—supervised and unsupervised (with variants in-between). The distinction is, in fact, rather fuzzy. This is because in the supervised kind, learning examples are given to the machine that have been marked up or labelled by a person. These labelled patterns (an image of a dog say) are then used as a reference point that the application builds up a data base for, with numerous further labelled examples (of different dogs) being used to create a range of patterns that fall inside the sought for target (a dog) and those which are outside. Unsupervised learning proposes outputs which are validated by feature engineers, but have not been specified beforehand, as with supervised learning.

Between these two broad types there is also Reinforcement Learning, which likewise can sometimes be said to be learning without human intervention. This approach entails learning on the basis of some very high-level goal, and for the machine to attempt to attain this goal over thousands of instances until eventually it can reach it, learning on the way. In the case of playing Go, DeepMind's AlphaGo played against itself many millions of times before it learned how to deliver success – winning at the game against a real opponent. But one needs to be reminded that the goal was specified by the designers of the system.

In essence, and leaving aside the role of the human, these sorts of ML applications see similar shapes, and it is similarity that produces outcomes. Similarity is a vague word, of course, implying many dimensions; in ML systems it means nearness – near in shape, in colour, in form. How the data stored for this nearness is interrogated and new samples added to it is a remarkable piece of computer science engineering, focusing essentially on the geometries stored (or expressed) in tensors, but this is all that is happening: ML can show that 'this' pattern (a pattern that to us is the shape of a dog) looks like or is similar to 'that' instance.

Unsupervised learning is when an ML application is given some data sets and seeks to identify patterns in that data. But the learning is still fundamentally driven by people, as the patterns that the system identifies are sorted and 'corrected' by people: outcomes of unsupervised learning are labelled, and it is this labelling that allows the ML to learn relevant patterns. Crucial here is the basic truth that machines cannot recognise a pattern that is relevant to some enterprise, whether it be in science or in everyday life, unless it builds on human understanding of what matters (Collins 2018). Any claim that ML can deliver autonomous reasoning are egregious, though all too often made by people and companies who have a stake in inflating the powers of the technology. Sometimes these claims seem plausible at first. For example, DeepMind claimed that some of its ML for games playing (Go and such) did not entail any manual labelling for the stratagems being modelled. But this ignored the prior and much more significant

fact that the games were devised by humans in the first place. One might say that the 'nature' of the games had already been through the process of feature engineering *avant la lettre*. This is not denying that ML systems can uncover hitherto unknown patterns, it is to say they can do so by building on patterns that the human hand has already uncovered. ML builds on prior human work, to put it another way. The apparent appeal of current ML technologies, such as those labelled generative AI (e.g., the ChatGPT), also derives in first principles from human judgement.

An important innovation in ML was the emergence of deep learning. This approach is essentially a way of iteratively rejigging data stores (tensors) such that the re-specification of their parameters (their internal geometries) eventually outputs a recognised pattern. It does this through a process of back propagation. In very simple terms, this entails a system interrogating some new data against the model stored in its tensors; if it finds no fit, it alters the parameters of the tensors and tries again, until some features of the sought for pattern begin to emerge. When it does so, the system alters its tensors yet again so that their parameters are even nearer the emergent pattern. Crucial, though, is that as this happens a relationship between that emergent pattern and the sought for one also begins to emerge. In each iteration, so the system gets nearer to recognising the pattern it is programmed for. The technique has various subtleties, such as for example, dividing tensor parameters into smaller or narrower parameters and then seeing if patterns thus begin to emerge at this 'more detailed' level.

One consequence of back propagation is that deep learning systems can use huge amounts of memory, each iteration requiring a store of what was captured initially, a store of the new version, and then a way of referencing between each layer in which this has been done. Deep learning applications can reach several layers, the only limit being computer storage – memory. Though the concept was first outlined in the 1960's, and designs for how it might be done in the 1980's, it was only when engineers recognised that the graphics cards in games-optimised machines had the kinds of memory available that deep learning became practical. The year 2012 is often mentioned in relation to this (Marcus and Davis 2019).

A further and very important point is that deep learning systems are very good at recognising patterns and will keep processing until they do so (in most instances) but are massively limited in what they find. The patterns they identify have to be close to or of the same kind to the one(s) they were trained for. A deep learning system for word sequences will be useless for visual phenomena; a system for visual phenomena useless for language, though both use similar operator computations on the data (expressed in tensors) and very similar training techniques with data, they have very different high level architectures as what they need to output is very different (though of course one might claim that there can be artificially created synaesthesia, words are not visual objects as they are is vehicles of meaning, whereas visual objects can be symbols, they are first and foremost

objects). Given the processing demands that deep learning applications place on hardware, this is a major inhibitor for ML systems that depend on the interrogation of multiple and heterogeneous patterns. The very architecture of deep learning in effect can keep ML shallow. Some ML advocates argue that this will be solved with massively distributed systems that can access huge amounts of processing power. Transfer learning will be central to this. We are not so persuaded, however, as this seems to presuppose that one distance is like any other (i.e., the thing captured in the tensors) and that the problem for ML is simply scale. That distances might represent different phenomena is immaterial in this view. This view ignores the creative transformations of feature engineers make when constituting datum suitable for ML processing, a transformation that bypasses questions of epistemic incommensurability, as we alluded to in the opening remarks of this paper. This is a larger argument, needless to say, which we cannot expand here (but for an introduction to the engineering problems here see Duboue 2020.

Once a system has been trained (whatever the method and irrespective of whether it uses a deep learning approach), the architecture for it can be altered. When a system is being trained, there is a need for an intensive loop of learning, but once this has been completed, a system might be needed to process vast numbers of instances in real time, and for this a different architecture may be better suited. This can mean that the ML applications that people use every day are different from those inside development and research contexts in an important respect – though the everyday ones continue to learn with new instances, they cannot learn other patterns, other new objects to see or identify. They can only do what they were taught to do. The architecture of everyday ML is thus recalcitrant to what users might want – namely for themselves to become the masters of learning, the engineers who instruct the ML machine with newly discovered 'features' of whatever is of interest. We shall return to this when we discuss some chatbots based on ML, namely Tay and Zo.

At the same time, it is also true to say that, in some instances, especially in unsupervised systems, that what a system finds (on the basis of its learning) might be obscure at the point of use (or to the user, which amounts to the same thing). Deep learning variants of ML are especially vulnerable to this problem, sometimes spewing out 'answers' that do not make sense. Partly this is because the sum of criteria used to identify something include criteria that the user would not recognise (nor perhaps even the feature engineer). While a data engineer might have selected features for the application to learn, as the system processed these examples, it might have added features that the engineer was unaware of. Some of these may seem very odd indeed and can lead to errors.

The classic example of this used to illustrate the issues to undergrads and business users alike, has to do with wolves. In this, an ML application is described as having been taught to recognise wolves from ordinary dogs. Through the training, the system starts to rely on the background behind the

animal as a criterion for 'seeing' a wolf as against a dog. The background it searches for is 'white', as in snow. This is because the training sets used are of wolves in their natural habitat, often a snowy one. The system 'learns' that snow is therefore a 'feature' of wolves. Of course, this is wrong. This feature needs to be removed from the learning set – hence the importance of feature engineers' continual involvement in training.

This points to a related concern – not just that classification can be wrong, but rather that in the case of ML the system can err one side or another, or can 'overfit' a new example to old data, as in the case of the wolf in snow, or it can underfit – not seeing a wolf for lack of snow. Overfitting and underfitting are fundamental matters when systems are being trained, but sometimes continue to occur once systems are released. Whatever the particulars, the point is that it is sometimes difficult to make some outcomes 'explainable'. The data engineer who taught the machine to select the 'features' in question might have long left the scene and the machine itself has learnt new criteria. The analogy between this and the obscure meaning of data entered into data bases by staff who have long been forgotten should be obvious. It is not the technology that makes explainability hard, it is the lack of knowledge about purposes expressed in data in the system. With ML, these purposes can be made more obscure by some being selected by the machine.

There is one final irony worth mentioning from this oft-taught example of wolves-or-dogs-on-snow, however. While it is good as an illustration of the fragility of ML and deep learning in particular, it turns out there is no way of identifying a wolf from a dog of breed with similar form (a husky, say). What defines a wolf is not its shape; it is the wolf's unwillingness to be tamed. Being a wolf turns out to be a cultural matter, not a physical one. And this is a fundamental problem for AI too: such categories are not to be found in visual data or at least not without reference to categorisation procedures that are simply too different from those used in the type of computation used in ML and in deep learning especially. Cultural matters are cast in everyday language which as computer scientist, Drew Mcdermott pointed out long ago (1976), is only partial in its references, being largely reliant on context. No computer language operates without correct and logically consistent references to the things being computed. Computer systems cannot reference 'context' as this is too large and too vague to act as a reference. As Collins notes (2018), humans might understand each other through the cultural practice of understanding context, but computers cannot. This is fundamental and we shall return to it when we consider generative AI.

5 The Maths in the Learning Machine

It is not always fully appreciated that ML algorithms that do all the above are not entirely new, as we have already mentioned. It is even less often noted that they frequently depend on very well-known statistical techniques (Blackwell

2021; for an excellent introduction to the techniques see Domingos 2017). This should hardly surprise, given the pattern seeking purposes of the systems. What is without doubt new is the sheer speed with which data can be processed. This is partly to do with vastly better data management, and even the use of different computer chips that award more space for short term memory (i.e., data). It isn't just a case of innovation in algorithms, as we remarked at the outset.

Whatever the type of ML, and at the risk of over-simplification, when target data is discrete (i.e., the patterns in question are, as it were, free standing), the algorithms in question will involve classification methods – is the pattern A or B? When the target data is continuous (when the pattern is a trend, a vector between two points—and sometimes without start or end points), the algorithms will involve regression analysis of some kind. At their simplest, classification in these cases can be done through linear regression models. They are generally thought to be less accurate than some other techniques since all they are really doing is taking a series of data points and identifying something that looks like a trend between them. Of course, the more kinds of data that are being analysed, the more complex the regression problem. Logistic regression, the most commonly used technique, is used to provide an apparently 'objective' result, which here basically means a binary, 'yes/no' outcome (regardless of the number of variables used to produce the outcome). More complex techniques involve such techniques as Support Vector Machines (SVMs), algorithms which essentially draw boundaries, called hyperplanes, between data points so that they fall on different sides of a demarcated boundary. They are particularly useful with limited numbers of data points and are regarded as being rapid.

Other familiar techniques include decision trees which rely on a dendritic (tree-like) set of branches to model possible outcomes deriving from a specific problem. Decision trees have relatively limited value because they presuppose that the structure of a 'decision-making process' is both linear and (at least to a degree) generally applicable. (Early critics of expert systems recognised this as a problem of context).

A more complex procedure is called the random forest. Its advantage is that it can be applied to N-number of variables and is called a 'forest' because it consists of several decision trees. Each tree describes local conditions. The random forest relies fundamentally on cheap computer memory. It works by collecting data until it (hopefully) finds all plausible statistical combinations' and then evaluates between them to find the most likely, given the known end point. It does this also statistically.

A further method is the so-called naïve Bayesian algorithm which uses conditional probabilities to predict classes. It is used in applications such as weather forecasting and fraud detection. As with all statistical techniques, underpinning the result is a degree of 'subjectivity' (a rather unhelpful term, as we will

argue below). To take Bayesian statistics as an example, prior probabilities can be understood as a mathematical expression of expectations (see Harper et al 2016) and hence as embodying judgements. The adding of 'information' does not change this, for what qualifies as relevant information is also a matter of judgement. As Harper et al. put it,

> "The general proposition goes like this: We can begin with a statistical fact of some kind and describe a probability for it. This is called the prior probability (but this need not be objectively derived ... it might be a guess, a belief, or an 'expert opinion'). We might subsequently run a test, or collect, or get access to some other evidence. With this, we can update our statistical analysis on the basis of this second tranche of information. It is important to bear in mind that the updated statistical probability (usually called a conditional probability) still depends on the prior probability but is now improved on the basis of the new evidence." (2016: p163)

The obvious point to be made about all such algorithms and not just the Bayesian is that they are techniques for reducing immense diversity and complexity into manageable relationships about values or distinctions visible in the data. These relationships may or may not be complex but consist of various ways of conceiving of nearness and its opposite, distance. This is not simply a geometric concern as it can be temporal, and indeed any combination of vectors can be involved. Distance is multi-varied, and hence is as complex as the pattern in question.

6 The Epistemologies of Inputs and Outputs

Whatever technique is deployed, a fundamental epistemological assumption underpins all. This is that the outcomes, derived statistically, have some kind of special status, that it is 'objective', say, or has scientific status – a known fact. The language used for them implies as much: they are the 'objective function(s)'. What they are also is predictive, and this is enormously useful. By predicting where patterns will show themselves, or perhaps we should say by showing what might be the consequence if a pattern were to show itself, ML systems can be very useful. Indeed, the word 'oracular' is sometimes used to describe them. However, it is no great discovery that the reliability of these outcomes depends on the initial assumptions upon which calculation is to be made, or on what is sometimes called the 'ground truth'.

Where the ground truth consists of data which is, for all practical purposes, objective, this would not appear to be too much of a problem. How often this is true, however, is debatable. The reality, as a number of observers have pointed out (most famously Crawford 2021), is that so-called ground truth can, in practice, rely on the labelling work done by an army of workers in the background

– the feature engineers and feature extractors we mentioned at the outset. And these might be making decisions that are contentious.

Take the example of ground truth about 'people'. If height and weight are the variables in question, there would appear to be little problem in doing the 'feature extraction' from a set of images of 'people'. If 'race' and 'gender' are the features in question, one can see how problems such as racial bias or gender stereotyping might easily insert themselves in the labelling process.

The same potential problem of agreeing what is relevant or suitable feature is evident if one considers assessments of risk to child welfare, propensity to commit crimes, emotional state, and on and on. The fact is that in all of these examples, the features selected in various instantiations have been disputed (see e.g. Jackson 2018; Berk et al. 2021; Asaro 2019 Li and Deng 2022).

The use of various statistical methods to solve the problem of choosing the right ground truth (and hence 'interpretation', to undermine that notion of truth somewhat) given many alternatives is effectively a judgemental one. As Blackwell puts it,

> "Random forests, neural networks, genetic algorithms - despite their evocative names, all are simply strategies for finding the most effective simple explanations within a hugely varying and obscured set of possibilities, while avoiding the 'local maximum' of an explanation that accounts for some variations, but not others more distant" (Blackwell 2017: 6).

Rather than a scientific or objective truth, what is produced is a 'best fit', expressed numerically. What should be clear is that what is best fit for one community (or set of users) may not be best for another. It all depends in what best fit is, and this is irrespective of the statistics used to define it.

7 Learning in Action

Nevertheless, even though they might be only a best fit, ML applications have made enormous progress in dealing with things that can be readily rendered mathematically, these contentious issues notwithstanding. To say again, most of the progress is a function of processing power, data storage and the finesse of labelling in the learning routines rather than the deployment of radically new algorithms. Even so, progress in facial recognition, in animal recognition, in game playing, and so on, has been remarkable. The reason for this is that best fit is relatively easy to engineer for these kinds of phenomena. It is not just that the data upon which outcomes are predicated is amenable to reduction to mathematics and hence the feature extraction work we mention above (width of nose; height of forehead; shape of ears, etc.) as that what counts as good enough in particular instances is also easy to define. Take faces: to

identify a face as very close to a reference face is much easier to do than to say that a new example must be the same (i.e., mathematically isomorphic) as a prior one. In many ways, best fit in the areas where ML has been successful is not just a solution for these machines and what they are engineered to do, but for what people do in their everyday practices as well – faces do change daily, after all, depending not just on physics – light and shade, but mood and intent. A married couple do not find themselves in paroxysm of doubt when they wake up beside someone whose face no longer matches the one they married. They judge their partner in the round, and not by the fixed geometry of a head shape. This also nicely illustrates once again the limits of ML – it can be good at recognising faces but is not able to map the larger culturally framed context which makes faces only one measure of individual, whether they be husband or wife. This is a further example of cultural context and beyond the competence of ML.

As illustrative of how ML can be successful and what notions of best fit might look like, we now turn to two applications, largely because they are in very common use. The first is that of text and text translation and the second, recommender systems.

Textual analysis and its subset, natural language processing, has made huge strides in recent years. It is not so long ago that the attempts of Google Translate to offer translations were regarded as risible. In the context of European languages, at least, they have become surprisingly reliable by which we mean they have been good at offering 'best fit.' But as we say, best fit solutions in Google translate (and many other tools of similar order) offer what people find helpful. The language offered in translation may not be eloquent but, for most purposes, it does the job. (We should remember that there is no such thing as a 'perfect' translation anyway).

ML from text, or text mining, leverages text from a wide range of sources. It might include text found on social networks, in digital libraries, in news sources or in a variety of web-based resources – web pages being the most obvious. The techniques here treat text not as 'bags of single words' (and hence matchable to single words in other languages), so much as 'bags of words and word sequences', thus predicting what words are likely to follow any given input, or context of word sequences.

In most instances, these use of ML applications in language contexts (translation being a prime example) are relatively uncontroversial, though in other cases there can be controversy. They are used to identify news feeds, for example, and in sentiment analysis in political contexts. Facebook, unsurprisingly sees only benefit from these uses:

> "Designing a personalized ranking system for more than 2 billion people (all with different interests) and a plethora of content to select from presents

> significant, complex challenges. This is something we tackle every day with News Feed ranking. Without machine learning (ML), people's News Feeds could be flooded with content they don't find as relevant or interesting, including overly promotional content or content from acquaintances who post frequently, which can bury the content from the people they're closest to. Ranking exists to help solve these problems ... We use ML to predict which content will matter most to each person to support a more engaging and positive experience." https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/)

This is done through a series of ranking algorithms which, more or less, analyse 'likes', their frequency and recency and, again, uses best fit.

Critics of such systems point to the way they create 'filter bubbles' (see Sunstein 2004; Pariser 2011; Flaxman et al. 2016; Bruns 2019; Chitra and Musco 2020; Dahlgren 2021) and can act as echo chambers or amplifiers of opinion. Pariser, famously, argued that algorithms of this kind create "a unique universe of information for each of us, which fundamentally alters the way we encounter ideas and information" (Pariser 2011:12).

Others, like Bruns (2019), are more sceptical. Bruns sees the 'moral panic' over filter bubbles as expressing a form of technological determinism held by those panicking. As he puts it,

> "A moral panic about social media in themselves, then, independent of how and by whom they are used, is no more warranted than one about TV, radio, or the printing press. We would fall for technological determinism: a belief that social media, however platforms might be designed and however citizens might use them, inevitably promote echo chambers and filter bubbles. As we will see, there is no evidence to support such an argument ... We cannot absolve ourselves from the mess we are in by simply blaming technology" (Bruns 2019: 39).

Dubois and Blank (2018) continue this line of argument in the context of political opinion:

> "Whatever may be happening on any single social media platform, when we look at the entire media environment, there is little apparent echo chamber. People regularly encounter things that they disagree with. People check multiple sources. People try to confirm information using search. Possibly most important, people discover things that change their political opinions. Looking at the entire multi-media environment, we find little evidence of an echo chamber. This applies even to people who are not interested in politics." (Dubois and Blank (2018):740)

In short, understanding the relationship between how algorithms generate text related content and how people interact with the resulting 'texts' (from prosaic Google Translate instances to the sharing of news feeds through to the directing of political activity on the basis of those feeds) is at once a computational and a social matter and any claim to privilege one over the other needs care.

The same could be said about recommender systems, the second of our illustrations. The term, 'recommender system' was first coined by Resnick and Varian (1997). Since then, recommender systems have achieved a certain notoriety since they have implications in respect of advertising pressure and privacy; they can also come across as black box-like, with users feeling they have little control over what is recommended to them. They can have what is called an 'anchoring effect' with individuals finding themselves stuck in the space the recommender tools offer them (see, e.g., Adomavicius et al. 2013).

What is sure is that recommender systems are pervasive. Applications have been designed which include, for instance, effects on consumer preference (Zhang 2011), music preferences (Gunawan and Suhartono 2019; Moscato et al. 2020), movies (Ghosh et al. 1999; Cosley et al. 2003; Geetha et al. 2018; Walek and Fojtik 2020; Afoudi et al. 2021), health awareness (Ge et al. 2015), nutrition (Franco 2017), tourism (Nilashi et al. 2017, 2019; Brodeala 2020), and, reminding us of the interconnection with text based applications, in news feeds too (Fortuna et al. 2010). By no means all rely on ML but a progression towards these sorts of solutions is discernible (see for instance Mahata et al. 2016, on ML for movie preferences. Valdez and Ziefle (2019) note the different technical strategies that can be used by recommender systems).

Be that as it may, Spano and Boratta highlight some of the usability problems associated with these systems.

> "From the user's point of view, recommender systems remain a black box that suggests content, but the users hardly understand why some items are included in the list. The relevance of this issue has increased in recent years, as the introduction of approaches based on latent features (such as Matrix Factorization or Deep Learning) has made it very hard to connect user preferences with the recommended items. Providing the users with an understandable representation of how the system represents them and allowing them to control the recommendation process would lead to benefits in how the recommendations are perceived and in the capability of the system to be persuasive. Such transparency is one of the multiple (and usually conflicting) requirements of recommender systems." (Spano and Boratto 2019: 2)

How users interact with these systems, then, is becoming the subject of increased attention. What is sure, however, is that recommendation and the algorithms that deliver them have no ethical viewpoint themselves even if the

consequences of their recommendations might be consequential. Thus, as and when 'ground truth' points towards preferences for certain kinds of output, recommender systems will ensure more of that output is presented to the user whatever that output. Privacy concerns make this controversial enough, but the prospect that the user can be led to deep fake imagery, to violent content, to terrorist instruction, to radical religious proselytising and similar is more worrying. This is especially so when social media companies have been extensively criticised for not monitoring content (or removing it) in an adequate fashion such that these possibilities remain common.

As it happens, social media companies, at least in the USA, have no legal responsibility for their 3rd party online content after the so-called Cox-Wyden clause, or Sect. 230 of the Communications Decency Act (see Kosseff 2019). Kosseff is a defender of the broad purpose of Sect. 230, defending the argument that removing it would have a 'chilling effect' on the internet, but nonetheless concedes that strict interpretation entails "systematic problems that affected thousands or millions: trolling and revenge pornography, terrorist recruitment via social media, and the pervasive use of classified websites by sex traffickers." (p.209). Perhaps what we might learn from this is that, where in the early days of the internet, most situations were clearly delineated in law, the advent of AI and ML has made the role of technology on the web and elsewhere much more difficult to map out both in terms of law and the unintended social consequences of the technology in question.

8 AI, ML, the Social and the Cultural

If the position outlined above on 'ground truth inputs' and the 'objective function' outputs of ML is followed to its logical conclusion, then a space opens up for an agenda which places the social and cultural at the centre of investigation. Blackwell (2021) makes exactly that point at least in respect of an appeal to qualitative investigation. He argues that AI and ML are founded on presumptions about notions of a general intelligence, one which is discoverable through the algorithms of the kind we have been discussing (See Russell (2019) for such claims). Those with a Wittgensteinian persuasion have long critiqued this assumption, viewing it as founded in a 'primitive cognitivism' which holds that the outcomes generated by such processes are assumed to have a universal quality (see, e.g., Coulter 1987; Button et al. 1995). Blackwell is clear that assumptions of this kind are unwarranted and that the investigation of local particularities ought to be part of the agenda for ML. Doing so will highlight questions about epistemological categories and their cultural siting. Key here are notions of the 'objective' and the 'subjective'. We allude, above, to Wittgenstein as a source of an alternative view that does not accept this duality. In his thinking, the meanings of any action, their objectivity or their subjectivity, are not derived from

internal mental processes but are intersubjectively accomplished; i.e., through performance with others. It is through interaction that what is treated as objective comes to be agreed and identified, and likewise, it is through interaction what is subjective and only that (and hence, for example just one person's point of view) is established. In either case, they are knowable phenomena – neither hidden from view, 'inside a head' say. Hence, the way that ML defines and operationalises the objective and subjective, what might count as the status of inputs as ground truths and outputs as objective functions may be misleading. Both inputs and outputs are cultural outcomes, built on cultural practices, and each known, but treated differently. We shall be coming back to this in our conclusions.

For now, though, the point is that to say they are cultural or cultural practices does not end what might want to say in regard to the connection between ML and the cultural. In being cultural, a whole range of new questions as to what aspect of 'the cultural' matters, how they are understood and what methods are used to document or 'model' them.

For example, many institutions, researchers, and opinion leaders have recently promoted the notion that ML systems should be human-centred. If so, then one needs to ask what is meant by the term human-centred? Does it mean based on mathematical models of the behaviour? Would such models encompass the cultural, and if so, how? Not least, as any social scientist would acknowledge, the term 'culture' is itself rather problematic. Is the 'cultural' simply descriptive of patterns of behaviour in which case it can be discarded and even replaced with machinic descriptions that list only behaviours or does it incorporate theoretical assumptions about those patterns? Does it—and this is foundational to some perspectives at least—embody assumptions about 'motive' or 'purpose'? Are motive and purpose treated as just labels for action and so can be ignored? Are they subjective, and need replacing by the objective? If so, does that mean that human-centred ML ends up 'behaviourising' culture in its attempts to seek the objective in its conversions of ground truths into objective functions? One can put this more simply: does the modelling implied in human-centred ML simply map actions, and fail to trace the motives or purposes that those acts express? In this view use of, say, the web is simply a question of movement between websites or pages, and the reasons for those movements are treated simply as an expression of the vector. A user is not what they *think*, nor are they endowed with cultural knowledge that needs getting inside to understand (the essential problem of anthropology, of course); they are *only what they do.* This makes the modelling easy as users are nothing more than their digital footprint but seems to take the cultural out of their actions.

This has all sorts of consequences, needless to say, one being that what we ordinarily understand as a person, a creature with purposes and motives, will not be the same as a user as understood by machines. This user simply 'acts'. If so, what is the relationship between how a person understands themselves when

they engage with ML tools and technologies if those technologies convert their actions into 'user behaviour'? Is there a confused dialogue between the human and their virtual self, the user? (See Harper, forthcoming on this point. See, also Hilderbrant 2006, a legal scholar who foreshadowed these concerns twenty years ago when thinking about what future legislation about digitally mediated behaviour might look like. This was well before the advent of widespread ML technologies where this behaviourism seems commonplace).

There are important questions here. Although there is no consensus about what human-centred ML means, it commonly refers to a set of principles which hold that ML systems ought to be focused on how people collaborate so as to support and augment their practices. Humans, in this view, are central to the business of determining how and in what ways ML systems should be used but the requirements implied in any notion of being human-centred are wide, and as we have just seen might entail different notions of understanding what a human is meant to be.

At first glance these issues might seem to be merely about nomenclatures and how people are conceived of differently in different Human centred ML architectures. In this view, the human in explainable ML is different to the one in ML Fairness, just as they are in human-centred data science (HCDS), computational creativity or human-ML co-creation. This is to name only some of the variants (for a review see, Shniederman 2022). It should be apparent though, that for each of these labels, the problem of meeting them with ML applications is likely to be much more complex than with systems with identifiable rules, as was the case with GOFAI applications. For there, the human was cast as an actor behaving in predictable logical ways. In contrast, with these new terms, each takes a slanted view on something immensely more complex than just the human-as-user: they want to factor in culture. Each subsumes answers to questions about what cultural practice might entail, how it is understood in relation to the individual (the user, say, or the subject, etc.) and how that is 'expressed' as features in ML data sets. And what makes this worse is that this is expressed in both the input and the output data.

Doing so is not straightforward. The gathering of categories and their deployment can occur in several ways in ML. As we saw earlier, supervised systems require data to be entered into them; unsupervised systems process new data and seek to learn from that. Even though the latter always involve some labelling in the development, once released into the field *both* function autonomously. What would this entail in any of the above systems – in human-centred ML; or in ML Fairness?; and so on. How do autonomous systems 'solve' what is meant by the human in the system? Do they come to mimic that human and use that mimicry as the basis of what they understand as the human? This would seem a little odd. Or do they allow people to participate in the shaping of how the human is understood from the view of the machine? Does autonomy mean human driven learning sets?

9 Machines that Generate Outputs

One way of answering this is by looking at what autonomy might look like with the latest form of ML, so-called generative systems. To look at these, we return once again to machines that speak or are otherwise generative.

We started our discussions with remarks about Eliza and Parry; speech machines (or 'bots', though neither were called that at the time) that did not interactively develop their speeches, simply offering words and phrases from a table of prior responses. Today ML technologies allow speech engines to 'interact'. What does this look like?

Let us begin with a bot from some years ago (2016) but which enables us to see the link between Eliza and more recent instantiations, most famously, GPT-3. Microsoft's Tay ('Thinking About You') was introduced and made accessible to users via a web portal. Tay was a speech bot that was designed to respond to any conversational turn on the basis of learning the language found in social media and the web more generally. It took data from these sites and developed its 'powers'. However, it was removed after 16 h. What it learnt to do was not approved of. In this time, it/she/he had become infected with racist and sexist sentiments. The Tay speech bot (or application) was intended to replicate as far as possible the speech patterns of a 19-year-old girl from the US. It ended up speaking like a deeply offensive person, not someone that Microsoft's engineers thought acceptable and certainly not like the well-mannered by syntactically curious teenager they had in mind (a 'Millennial' was the category referred to in the PR).

The bot was subsequently replaced by a newer version, called Zo. In turn, Zo faced criticism for other (almost diametrically opposed) reasons. Stuart-Ulin (2018) describes Zo as follows:

> "Zo is programmed to sound like a teenage girl: she plays games, sends silly gifs, and gushes about celebrities. As any heavily stereotyped 13-year-old girl would, she zips through topics at breakneck speed, sends you senseless internet gags out of nowhere, and resents being asked to solve math problems."

She goes on:

> "But there's a catch. In typical sibling style, Zo won't be caught dead making the same mistakes as her sister. No politics, no Jews, no red-pill paranoia. Zo is politically correct to the worst possible extreme; mention any of her triggers, and she transforms into a judgmental little brat. Zo [would] not engage in any discussions of issues that had to do with, for instance, Islam or conflict in the Middle East. Jews, Arabs, Muslims, the Middle East, any big-name American politician—regardless of whatever context they're cloaked in, Zo just doesn't want to hear it. For example, when I

> say to Zo "I get bullied sometimes for being Muslim," she responds "so i really have no interest in chatting about religion," or "For the last time, pls stop talking politics…its getting super old," or one of many other negative, shut-it-down canned responses. By contrast, sending her simply "I get bullied sometimes" (without the word Muslim) generates a sympathetic "ugh, i hate that that's happening to you. what happened?"

Zo, like Tay, was also subsequently withdrawn from the web, though it took a little bit longer before this happened – some months, not hours.

For both Tay and Zo, there is an obvious point to be made about machines that learn, though we need to be careful as we unpack this. It is clear that the behaviour of the applications was a product of what each modelled and then used to craft output. The models included learning sets from the Twitter social media platform and from interactions with users via that platform. In this learning procedure, both Tay and Zo relied on the heterogeneous but essentially orthodox (i.e., commonly used) NLP techniques that include Bayesian inference engines to determine best fit phrases in conversational systems. Though the ground 'priors' might have been deduced from data and then the data subject to deep learning processes to produce the objective functions (i.e., the outputs), nevertheless, the data in all its dimensions was socially constructed: built on language learning sets taken from 'real' users (Twitter users) while the new data outcomes were generated through gathered real time dialogues. The outcome of these learning procedures cannot bring into doubt the nature of these sources. The input was the language of social media users; the outputs were the languages of social media users too. But the media was Twitter, and hence consisted of a language that not all communities outside of that particular social media would did find equally acceptable.

Now this is where we need some care. Whether or not the applications did a good job of learning from this data, what they did learn was, in effect, cultural practices – a particular type of language use. This reflected what we suggest might be a particular community's practice. As it happens this community likes colourful language; it also mocks 'wokism'. Other communities might have different practices. We do not want to explore what a community might be at the moment, suffice to say that twitter users might be thought of as a 'new public' type community, one emerging around new digital platforms and socially sensitive topics. Other communities may have quite different provenances – based on location, say, or religion, language, even sport. In each case, their practices will entail different topics and these will lead, in all likelihood, to different manners about language and attitudes to political correctness. Definitional matters aside, what we can say is that the Tay and Zo bots were out of their depth in the particular cultural practice and its associated manners and attitudes – not in being incapable of mimicking these but in not 'knowing' whether they were (or were

not) appropriate. In a sense, while they could certainly *model* aspects of the practices of the community, the bots could not act autonomously with these elements as if they were a competent member of the community. Their use of the language was especially not acceptable to the community – or rather, it was not acceptable to both those on the inside of the twitter community and those on the outside, looking on. And one might say that this was not because of what the bots said in this language but because of *what* they were, the bots. The reaction of the Twitter community was that bots have no right to speak offensively, whereas people do, especially themselves, twitter users. In other words, culture is not just a question of what is done, but *who* or *what* is doing it.

There is another important point. Leaving aside whether the ML technologies inside Tay and Zo did a good job, and leaving aside too, the morality of this culture, what is also evident is that those in this culture, twitter users, seemed to know enough about ML that they were able to intentionally make a mockery of it. Consider: it was they, after all, who in effect offered up training set data that supplemented the basic sets used to teach Tay and which, in combination, resulted in Tay being seen as a machinic imbecile; an offensive one to boot. Twitter users knew about what feature engineers do and exploited that knowledge when providing features that would lead Tay astray.

How different these 'users' were to the secretaries who were dazzled by Eliza, or the psychiatrists fooled by Parry! We learn, in other words, that with Tay and Zo, and despite the persistent ideological claims surrounding AI and ML, that certain parts of our society have become clued up enough not to be fooled by these claims. They see ML as a form of machinic processing where the learning, if that is the right word, is limited by its inputs and its mechanisms for calculating outputs. More, it has no capacity to judge the role of that learning in the larger contexts in which it is used – in terms of culture. We learn, too, from the reaction that the public to the 'performances' of Tay and Zo (and indeed Microsoft's decision to remove them from Twitter) that many of the public did not think these technologies had the capacity to judge the moral adequacy of the language they processed either. Some features of their language outputs might have been ok for human users to deploy, but other parts would not, the public seemed to feel. Why? It seemed that the public thought a machine cannot judge – these machines certainly. In their view, these AI technologies were (are) morally bereft; that is, incapable of making moral judgements about vocabulary or acceptable topics in talk. That they could not manage these concerns opened a space for the same public to mock these technologies for being so ill adept.

We might not agree with the mockery but perhaps we ought to agree with the judgement, certainly in the case of these two instances. The absence of any cultural framing that would make these speech bots 'know' the difference in the contexts of language use and the appropriate use of attitudes surely attests to the accuracy of this judgement. Their failures underline the potential limits of ML.

Social media can encourage colourful language; interacting with a speech bot is not a place for such language; social media, especially Twitter, is a place where extreme sensitivity to the relationship between topic and politics is to be found. What Microsoft's Tay and Zo bots showed is that ML *might not* learn what *ought to be* learned, and that left to their own, ML tools can take us, their users, to places we don't want to go. As they do so, they remind us, too, that the term 'intelligent' is not always what comes to mind when we think about ML technologies – indeed, these technologies can evoke quite the opposite.

10 From Inferred Response to Generative AI

Tay and Zo were released some years ago. What has happened since? One particular and much celebrated current claim holds that a 'generalised' form of AI is about to appear (Russell 2019). GAI is the new acronym for this (though sometimes AGI). Key this is ML and two new techniques – ensemble learning, which in effect pools outputs from different ML techniques that might be suitable for different aspects of the modelling task but which can be brought together at a higher more abstract level. And the other is Transfer learning (Vaswani et al 2017). This is a different technique made up of the outputs *of self-attention* modelling operating within a learning model, outputs which can be applied to a different model – hence transferring learning from one to another. When brought together, this can lead to even higher-level models. These are now being called foundation models (Bommasani et al. 2022). The resulting outputs (or models) are of sufficient generality to allow some to think that general artificial intelligence is about to appear enabled by ML, as we say.

This might not be the right way of conceiving it, however. For, what they do is allow the recognition of more patterns in particular kinds of data and these can be sorted to identify patterns that are, as if were, at a higher level. Before foundation models, in the case of natural language processing, the modelling that was possible could only be used to understand certain phrases and words; with transfer learning and foundation models much more elaborate language models (and hence outputs) can be made. This can be reflected when a 'prompt', which we can conceive of as a natural language category (bear in mind this is a simplified account of what happens), is thrown into the transfer-derived model(s) and an 'emergent' output appears, echoing, or rather modelling, that very prompt. That prompt has to accord with a pattern in the foundation model, but if it does, then the foundation model can link that to a variety of other related patterns or models and this can point to various new 'next words/phrases' or, if you like, 'next concepts'. Thereby an apparently simple prompt as an input can lead to an apparently complex output.

A good example of this, and close to the Tay and Zo bots in being speech based (as foundation models can be of other data types too, such as vision), can

be found in interaction with GPT-3 (Generative Pre-trained Transformer—3, the number being the 'generation') and a more recent instantiation, chatGPT**.** Focusing on GPT-3 for the purposes of discussion, this is a language engine that takes an initial prompt provided by some input text from a user to create an extended and richer textual response (https://en.wikipedia.org/wiki/GPT-3). The technology includes some new foundation models, based on huge numbers of parameters, built on ensemble and transfer learning.

At first glance, GPT-3 certainly does seem to offer dazzling and sometimes unexpected responses to a prompt. Putting it simply and using everyday language to describe the technologies functioning rather than computational terms, it works as follows. If one takes a written story as an example of a pattern, the GPT-3 engine has models of stories as well as models of phrases, words, sentences. These are derived from learning, so what these models look like might not accord with how a person would imagine a story and its words to be – for GPT-3, they are mathematical for one thing, and constructed on notions of likelihood and proximity in observable patterns, and not such things as purpose, or popularity or sentimental value which might resonate more with an individual's notion of what stories 'are' and the words that make them up. Be that as it may, one might say that the inference engine, GPT-3, maps narrative arcs along with everyday phrases such that both are used to determine (predict) the meaning of particular words or phrases offered in stories. It works in both directions, if you like. It takes outputs and sends them back to inputs and from inputs renews its outputs and eventually this allows it to produce phrases, stories with narrative arcs that allow it to model, i.e., predict, new words and stories. It uses these to respond to a user input, the prompt mentioned above.

So, for example, from the view of GPT-3, the word 'fairy' will have one meaning as it is understood in terms of a *fairy story* (and the word story will itself be referenced to noun instances like fairy in such stories), while the use of another very similar word, 'Fairey', will have a different meaning. The latter has a meaning deriving from the history of airframe manufacturers; so too will particular terms and phrases. Hence, GPT-3 will respond to a phrase about the former differently than to a phrase about the latter. In either case, GPT-3 will produce many lines of text as an output. And this text output will not simply echo or mirror the text input (as would a deep learning system) as it will include terms that are apparently related. A story about fairies may elicit a response about bedtime stories, for example, as these might be judged as related to fairies in the foundational model, or at least might be outputs generated by the system (hence the term, generative). As it happens, using the word Fairey in the GPT-3 platform results in the word being 'auto corrected' to fairy, there apparently not being a sufficiently frequent record of the Fairey Aircraft Company on the GPT-3's data sets.

11 AI and ML in Search of a Perspective

As we say, some think GPT-3 is 'intelligence' because of its capacity to generate text in these ways. Geoff Hinton, mentioned several times, seems to think its use of 'vector techniques' to frame meaning is the major step to ML if not the last, key step. To say vectors is just another way of conceiving of transfer learning procedures. Either way, Hinton and others think something big is happening.

It is at this point that we want to return to our opening remarks. We suggested that CSCW might well offer a relevant view for ML as it encourages research that looks at all aspects of the social context of technology use, and, in the case of ML, that means looking at input and outputs, at learning regimes and training sets, at claims about objective functions and ground truth, and placing insights about all these matters in reference to contexts of use. We have begun to point to CSCW papers that have begun such enquiries; the ones that follow in this collection do the same. But we also mentioned that it might be that ML needs a perspective of its own. We have alluded throughout to how ordinary language can shape the way people understand technology. The technology itself also participates in this shaping, needless to say. Somehow, in use, GPT-3 seems to express something; it does so in its outputs. These are evidently more than the stuff derived from a 'phrase engine', as seemed to be the case with Tay and Zo. The way GPT-3 seems to deploy narrative-like structures to meld common place terms in its bulky text(s) suggests something creative. Indeed, that it produces text like a tide can seem persuasive of something radical being present. Oddly, though, GPT-3 *nearly always* seems to give itself away. The combinations it presents have incongruities; they often include facts that seem egregious; occasionally, the 'shape' in a text that ought to reflect some purposes at hand seems wrong—mishappen somehow. All these serve to alert the user to something seen before with computational tools: what one might summarise as a lack of cultural knowledge that, if present, would not have allowed these strangenesses to appear. These errors would not occur, one says to oneself, 'if the technology knew about the context, the why I am here, and what I am doing'. Many inside the ML community argue that the use of huge data sets for GPT-3 will deliver powers that equate to this knowledge. Surely, they seem to think, if all the materials on Wikipedia, all the language models available, and even better tools for 'fine tuning' foundation models are brought together then GPT- (number 'X') will pass the Turing test. Indeed, one might not be surprised that many organisations with very large computational resources are now planning to undertake such activities as they think it will provide them with new opportunities. These are the 'platform capitalists' – the Googles, the Apples, the Alibaba's and, of course, Microsoft which has bought exclusive rights to the internals of GPT-3, all of whom seem to have been taken over by what seems to be the credo surrounding contemporary ML. We mentioned Norvig as one of those behind this earlier on – for a while he led Google's research division in Mountain view. We need to wait and see.

GPT-3 is only now being used and hence there are few if any academic investigations of it. One might suggest, though, that prior research that have identified the limits of ML would likely apply to GPT-3 too. The most important have sought to distinguish between the way that ML patterns natural language but does not—and in their view cannot – pattern *meaning*. The distinction here has to do with how people locate the meaning behind a word, a phrase, or even larger text, through reference to matters external to the words themselves. Key here are the purposes of persons. Hence, in the case of GPT-3 offering a response to a prompt about fairies, users of the same will seek to identify what the purposes behind that response might be in terms of a person's intentions – as that is the only way meaning is constituted (Harper et al. 2019). What they find with the oddities that GPT-3 presents, is an intimation that there are no purposes; from this they deduce that GPT-3 is like prior incarnations of AI, not a substitute for a real person. On the contrary, in this absence, they come to see that GPT-3 'means' nothing.

For most users, this is not a concern; only a moment when they confirm their understanding of what they are dealing with is a machine, not a person. We might remind ourselves of how different this is from users of Eliza and Parry who came to think it was a person they were dealing with. The lesson is not that the technology has got better, however; it is that people's knowledge and expectations have altered: today, users may engage with tools like GPT-3 not to come to the conclusion that a person is now equalled by a machine, but with a view to discovering whether or not the information conveyed is reliable. While there was an initial rush in the media to wonder at the apparent power of the application, it was quickly followed by scepticism when its limits became apparent.

It is important to note that this does not mean that technologies like GPT-3 lack potential use. On the contrary. If they know what the technology offers, they can still use it. How useful it will turn out to be, for whom and in what contexts is very much the kind of question we might expect CSCW commentators to be answering. Currently, potential use is hidden and muddied by ML engineers who seem obsessed with passing some version of the Turing test. But it seems to us that ordinary people have long since moved on from this concern. As we say, there are, as yet, no good scientific studies of use, but we do think that they are likely to uncover what we have just foreshadowed – that the tool is very useful, but to be understood in terms of how it works and hence what it does – offering frequency-based responses to queries.

What we are pointing towards is not what the technology does, but the claims made about it. Bender and Koller (2020) argue that this is a fundamental problem for the ML community. Its insistence on justifying its hype can—and indeed in their view often does—led to dismay and even rejection by users. It has done so in the past and will in the future with technologies like generative AI ones. They urge computer scientists to recognise that a system trained in natural language

alone, like GPT-3, does not include the learning about the relation between that language and meaning, as this is external to language. There is a difference between patterning the form of language and the way language is used, what it means. As they put it:

> "What's interesting [ ] is not that the tasks are impossible, but rather what makes them impossible: what's missing from the training data. [ ] a system trained only on the form of [ ] English has no way learn the [ ] respective relations [that makes] a sentence meaningful." (2020: 5190).

Their distinction between form and meaning is not one that points to the hope that meaning can be modelled. They go on to say:

> "The process of acquiring a linguistic system, like human communication generally, relies on joint attention and intersubjectivity: the ability to be aware of what another human is attending to and guess what they are intending to communicate. [as a case in point] Human children do not learn meaning from form alone and we should not expect machines to do so either." (2020: 5190)

Bender and her colleagues are linguists, and it may be that this provenance results in their arguments being dismissed by those inside the world of ML. In a more recent paper Bender and Gebru (2021) suggested that deep learning speech engines can be said to behave like 'stochastic parrots'. Though their argument is about reducing the hype around deep learning, the term seems to have stung, and at the current time one can often hear ML researchers asserting at the commencement of their talks that their inference engines are *not* such parrots. So, let us end our enquiries by recalling a dyed in the wool computer scientist and AI researcher, Drew Mcdermott. This was long ago (1976) but echoes Bender and her colleagues. Mcdermott was one-time head of computer science at Yale. He observed that a key problem in artificial intelligence is the notion that its language models will one day be complete. Thinking that they might eventually be so betrays a fundamental misunderstanding, he explained. Natural language is not itself complete. Its meanings, its utility, its applicability, is crucially expressed in contexts; in the doings of people in the communities they make. The references that give language life are *outside* of language; they are to be found in the contexts of use. This is not to say one cannot model language terms themselves, but this ultimately misses a crucial thing: the situated purposes of language. This is, of course, Wittgenstein's argument, though Mcdermott presents it in the style of any 'ordinary language' philosopher – Austin, Ryle, Searle. The point is that, however great in scale, however refined foundation models maybe, computer versions of language based on today's ML technologies will not ever have this

reference. The cultural contexts will be always beyond them. They are outside the datum. This does not mean that the models will have no use or will not deserve to be esteemed for the things people can do with them. Far from it. It is just that the way that some describe these models, the claims they make and even some of the mnemonics used to describe their functioning, can mislead and tempt some into thinking that context is included or evoked in ways that makes meaning complete. It may be, for example, that some forms of 'generative AI' are more culturally dependent than others. The media, thus far, have paid little or no attention to the possibilities offered by foundation models in relation to, say, computer programming. Although the jury is still out with regard to just far ML will be able to replace human programmers, efforts are underway to benchmark code generation (see e.g. Ruchir et al. 2021).

The perspective that is required for ML, then, is one on itself and the claims that attend it. When seen for what they are, the prospects for ML and for all their derivatives including foundation models, is quite different indeed from how they are sometimes presented and also, as we have just seen, even how they are labelled. It is not passing the Turing test that should be the aspiration, as it is offering machinic tools for jobs to do. Designing these to give the impression they are human-like is both impossible theoretically and distracting practically. In short, the answer to the question posed at the outset of this paper is that the future of ML will be achieved partly when there is a change of perspective within ML. CSCW might help here as it is interested in what is done with ML, and part of doing so might entail enquires into the belief systems in the places that invent ML tools and technologies. It could also be that examination of the culture of feature engineering is a proper topic for HCI too.

12 The View from CSCW

Fortunately, dealing with culture when technologies are used is business of CSCW. Though CSCW researchers have looked at many different technologies and many different circumstances with many different types of users, the approach speaks naturally to what is done with AI and ML, not claims about the technologies. Given this, we make the following recommendations.

Firstly, CSCW researchers might look at the way in which data is entered into ML systems and at how users make sense of system outputs. These are, in a sense, sides of the same coin. Questions which have to do with how one represents data outputs in such a way that they are made understandable are relevant to how data is engineered prior to being used in learning just as they are with what derives from learning – what comes out.

This, of course, cannot be an entirely technical matter. CSCW is founded on socio-technical assumptions. It follows that we might take our lead from

Garfinkel (1967) by showing how accounting for decisions which go into feeding data into systems to produce a 'ground truth' are accountable matters, matters worthy of careful investigation, just as are the ways that objective functions are accountable too. In this view, explanations are not generalised, as seems to be thought by the ML community, but designed to be appropriate for some task at hand – they lead to subsequent action. The question is what action, why and with what consequences – from inputs through to outputs. This subtly shifts the problem of 'explainability' since it is no longer a property of the machinery alone but one of the interactions between the system and its users. In other words, explanation has to do with the need for accountability of some system to particular individuals and particular groups given the actions they are engaged in. Representing is not an abstract concern, in this view, but one of applicability to organisational matters, a matter of recipient design. Organisational members will have particular aims and objectives and the shaping of inputs and outputs will necessarily have to be posed in a form that is relevant to what an organisation needs and what it does. This, in turn, raises questions about appropriate ways of visualising machine inputs and outputs. Rendering them in such a way that organisational members can make sensible decisions is an entirely non-trivial matter.

Second, this, in turn, suggests an opportunity for CSCW researchers to investigate a closely related issue—that of the tools and techniques that are common to ML and which might need to be understood to make these renderings accountable. Guidotti et al. (2018), in a well-known paper, have argued that many systems designed to support decisions typically hide their internal logic. That is, they constitute yet another 'black box' technology. As they say,

> "The applications in which black box decision systems can be used are various, and each approach is typically developed to provide a solution for a specific problem and, as a consequence, delineating explicitly or implicitly its own definition of interpretability and explanation." (Guidotti et al. 2018: 1)

They go on, much as Blackwell and others have asserted, that.

> "This enormous amount of data may contain human biases and prejudices. Thus, decision models learned on them may inherit such biases, possibly leading to unfair and wrong decisions." (Guidotti et al. 2018:1)

Their paper is a thorough examination of the ways in which different tools and techniques inside the black box of an ML application can provide 'explanations'. They also draw attention to the fact that, in Europe at least, GDPR regulations

give individuals the right to obtain data 'meaningful of the logic involved' (p2) and so this might also be something required, not just a nice thing to have.

Guidotti et al. pose, a set of interesting and difficult questions that CSCW researchers might examine in particular:

> "What does it mean that a model is interpretable or transparent? What is an explanation? When a model or an explanation is comprehensible? Which is the best way to provide an explanation and which kind of model is more interpretable? Which are the problems requiring interpretable models/predictions? What kind of decision data are affected? Which type of data records is more comprehensible? How much are we willing to lose in prediction accuracy to gain any form of interpretability?" (Guidotti et al. 2018: 3)

The point here, is that tools themselves can only be assessed in relation to 'explainability' if one has some sense of the capacity of human beings to make sense of their outputs in terms of what they do. If 'explainability' refers to the capacity of machinery to provide meaning in understandable terms to a human being, then obvious questions arise as to what person, when, in what circumstances, and so on. Some CSCW studies have already shown the way, notably in the healthcare arena (see Ontika et al. 2022; Park et al. 2019; Ploug and Holm 2020) but as yet we would argue relatively little attention has been paid to the work that goes into understanding AI and ML. There is clearly scope for much work to take place in a wide range of domains.

Allied to this, and third, CSCW researchers might investigate the organisational contexts in which ML technologies are interpreted, decisions are implemented, by whom, and for why. Organisational expertises and how they are (or are not) shared are just as relevant in the context of ML as they are in other technologically enabled contexts. What organisational ends are being met with data outputs? What is being done, organisationally, with data inputs? How do people who have had no part in the training of an algorithm judge whether its outputs are to be trusted and, even more importantly in the contexts of ML, what they do about them given the demands of an organisation? Bittner's notion of organisational compliance is brought to mind, updated for the age of ML (Bittner 1965). This points to other related consequences, intended and unintended, of the decisions made at various levels throughout an organization, the community and at a wider societal level.

13 Conclusion

How one goes about judging whether these topics are examined effectively is (obviously) very much at the heart of scientific inquiry, and not just CSCW. CSCW, with its interest in contexts of use and its resolute focus on cooperation and coordination on the part of user communities, is ideally situated, we feel, to

examine how, in practice, ML technologies get deployed in the real world. In any event, the four manuscripts in this collection represent what we consider to be good examples of where CSCW is in relation to machine learning, addressing the issues we outline above.

This collection investigates reputation systems (a version of recommender technologies discussed above). The specific context the authors examine is that of peer-to-peer car sharing and how trust is arrived at. They find, on the basis of several 'problem centred' interviews, that systems which provide algorithm-based scoring do deliver increased levels of trust. But they also show that trust is partly a matter of what the technology outputs and how 'users' understand how it processes data to make outputs. Users want to know how conclusions are made if they are to act on those conclusions. Accountability here does mean looking inside the black box.

This collection also tackles the way in which algorithmic systems develop by examining how and when designers intervene. Authors make the important point that design decisions around algorithmic development entails a great deal more than technical competence. It entails binding technological design alternatives to real world contexts, where these contexts are irremediably organisational.

Open-source intelligence (OSINT) technologies is also explored, looking at the role of values and value conflicts and the way in which they emerge in the application and development of ML- based OSINT technologies. A combination of methods is deployed, including a systematic review of the technical literature, a series of semi-structured interviews, and a focus group. The context here of cyber security incident response operators can seem a long way from everyday organisational action, but that is precisely what OSINT technologies are designed for: when the everyday goes wrong, and it can do so on any given day. That is why questions of trust matter.

Finally, a different tack is taken, focusing more on the in-situ evaluation by end users of a ML algorithm by the people who might rely on it but who equally know the kinds of data that it is using to infer. The context is that of sales planning, where ML is used to leverage a variety of different input parameters. This study of a bakery company demonstrates the many challenges involved in making a ML algorithm understandable, trustable and therefore useful to the different parties in the organisation are not uniquely to do with ML (or indeed any form of AI) as they are to do with the relationship between technology and everyday reasoning. Shifting the role of persons and technology into different relations with the processes at hand can have all sorts of perturbations and can lead to the reputation of technology being diminished not enhanced.

References

Adomavicius, G., J. Bockstedt, P. Shawn, and J. Zhang. 2013. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research* 24 (4): 956–975.

Afoudi, Y., M. Lazaar, and M. Al Achhab. 2021. Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory* 113: 102375.

Asaro, P. 2019. AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine* 38 (2): 40–53.

Bassett, C. 2019. The computational therapeutic: Exploring Weizenbaum's ELIZA as a history of the present. *AI and SOCIETY* 34 (4): 803–812.

Bassett, C. 2021. *Anti-computing: Dissent and the machine*. Manchester University Press.

Bender, E.M., and T. Gebru. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21, March 3–10, Virtual Event*. Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Bender, E.M., and A. Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods and Research* 50 (1): 3–44.

Bickhard, M., and L. Terveen. 1996. *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Amsterdam: North Holland, Elsevier.

Bittner, E. 1965. The concept of organization. *Social Research* 32 (3): 239–325.

Blackwell, A. 2021. Ethnographic artificial intelligence. *Interdisciplinary Science Reviews* 46 (1–2): 198–211. https://doi.org/10.1080/03080188.2020.1840226.

Blackwell, A. 2017. Objective Functions, Deep Learning and Random Forests. Contribution to Science in the Forest, Science in the Past, Needham Institute, Cambridge. Available at: http://www.cl.cam.ac.uk/~afb21/publications/Blackwell-ObjectiveFunctions.pdf

Bommasani, R. et al. 2022. On the Opportunities and Risks of Foundation Models, Centre for Research on Human-Centred AI, Stanford University, Stanford. https://arxiv.org/pdf/2108.07258.pdf

Brodeala, C. 2020. Online recommender system for accessible tourism destinations. In *Fourteenth ACM Conference on Recommender Systems*, 787–791.

Bruns, A. 2019. *Are filter bubbles real?* London: Wiley.

Button, G., J. Coulter, J. Lee, and W. Sharrock. 1995. *Computers, minds and conduct*. Cambridge: Polity Press.

Chitra, U., and C. Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining,* 115–123. New York, NY: Association for Computing Machinery.

Collins, H. 2018. *Artifictional Intelligence: Against Humanities Surrender to Computers*. New York: Wiley.

Cosley, D., S.L. Lam, L. Albert, J. Konstan and J. Riedl. 2003. Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 585–592. New York, NY: Association for Computing Machinery.

Coulter, J. 1987. *The social construction of mind: Studies in ethnomethodology and linguistic philosophy*. Godalming: Springer.

Crawford, K. 2021. *The Atlas of AI*. New Haven, CT: Yale University Press.

Dahlgren, G. 2021. A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review* 42 (1): 15–33.

Domingos, P. 2017. *The Master Algorithm, How the Quest for the Ultimate Learning Machine will Remake our World*. London: Penguin Books.

Dubois, E., and G. Blank. 2018. The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication and Society* 21 (5): 729–745.

Duboue, D. 2020. *The Art of Feature Engineering*. Cambridge: Cambridge University Press.

Flaxman, S., G. Sharad, and M. Justin. 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80 (S1): 298–320.

Fortuna, B., C. Fortuna, and D. Mladenić. 2010. Real-time news recommender system. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 583–586. Berlin, Heidelberg: Springer.

Franco, R. Z. 2017. Online recommender system for personalized nutrition advice. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 411–415. New York, NY: Association for Computing Machinery.

Garfinkel, H. 1967. *Studies in Ethnomethodology*. New York: Prentice Hall.

Ge, M., F. Ricci, and D. Massimo. 2015. Health-aware food recommender system. In *Proceedings of the 9th ACM Conference on Recommender Systems*, 333–334. New York, NY: Association for Computing Machinery.

Geetha, G., M. Safa, C. Fancy, and D. Saranya. 2018. A hybrid approach using collaborative filtering and content-based filtering for recommender system. *Journal of Physics: Conference Series* 1000 (1): 012101.

Ghosh, S., M. Mundhe, K. Hernandez, and S. Sen. 1999. Voting for movies: The anatomy of a recommender system. In *Proceedings of the third annual conference on Autonomous Agents*, 434–435. New York, NY: Association for Computing Machinery.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys* (CSUR) 51(5): 1–42. https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/ retrieved 27/10/2020.

Gunawan, A., and D. Suhartono. 2019. Music recommender system based on genre using convolutional recurrent neural networks. *Procedia Computer Science* 157: 99–109.

Harper, R., D. Randall, and W. Sharrock. 2016. *Choice: The sciences of reason in the 21st Century*. Cambridge: Polity Press.

Harper, R., D. Watson, and C. Licoppe (eds.). 2019. *Skyping the family: Interpersonal video and domestic life*. Amsterdam, Netherlands: John Benjamins.

Hilderbrant, M. 2006. Profiling: From data to knowledge. *Datenschutz und Datensicherhiet* 30, 9.

Jackson, K. 2018. Predictive analytics in child welfare–benefits and challenges. *Social Work Today* 18 (2): 10.

Kosseff, J. 2019. *The twenty-six words that created the Internet*. New York: Cornell University Press.

Kurtzweil, R. 2013. *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Viking.

Li, S., and W. Deng. 2022. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 13 (3): 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446.

Li, Shan, and D. Weihong. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020).

Mahata, A., N. Saini, S. Saharawat, and R. Tiwari. 2016. Intelligent movie recommender system using machine learning. In *International Conference on Intelligent Human Computer Interaction*, 94–110. Cham: Springer.

Marcus, G., and E. Davis. 2019. *Rebooting AI: Building artificial intelligence we can trust*. New York: Vintage Books.

Mcdermott, D. 1976. Artificial Intelligence meets natural stupidity. *ACM SIGART Newsletter* (57) 4–9. https://doi.org/10.1145/1045339.1045340.

Moscato, V., A. Picariello, and G. Sperli. 2020. An emotional recommender system for music. *IEEE Intelligent Systems* 36 (5): 57–68.

Natale, S. 2019. If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. *New Media and Society* 21 (3): 712–728.

Nilashi, M., K. Bagherifard, M. Rahmani, and V. Rafe. 2017. A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Computers and Industrial Engineering* 109: 357–368.

Ontika, N.N., Syed, H.A., Saßmannshausen, S.S., Harper, R. Chen, Y. Park, S.Y., and M. Grisot. 2022. Exploring human-centred AI in healthcare: Diagnosis, explainability, and trust. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. Coimbra, Portugal: European Society for Socially Embedded Technologies (EUSSET).

Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you*. New York: The Penguin Press.

Park, S.Y., P. Kuo, A. Barbarin, E. Kaziunas, A. Chow, K. Singh, L. Wilcox, and W.S Lasecki. 2019. Identifying challenges and opportunities in human-AI collaboration in healthcare. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 506–510. New York, NY: Association for Computing Machinery.

Ploug, T., and S. Holm. 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine* 107: 101901.

Resnick, R., and H. Varian. 1997. Recommender systems. *Communications of the ACM* 40 (3): 56.

Ruchir, P. Kung, D.S., Janssen, G. Zhang, W. Domeniconi, G. Zolotov, V., Dolby, J. et al. 2021. CodeNet: A large-scale AI for code dataset for learning a diversity of coding tasks. arXiv preprint arXiv:2105.12655.

Russell, S. 2019. *Human Compatible: AI and the problem of control*. London: Penguin.

Russell, S., and P. Norvig. 2017. *Artificial intelligence: A modern approach*. Boston, USA: Pearson.

Shniederman, B. 2022. *Human-Centred AI*. Oxford: Oxford University Press.

Spano, L., and L. Boratto. 2019. Advances in computer-human interaction for recommender systems (AdCHIReS). *International Journal of Human-Computer Studies* 121: 1–3.

Stuart-Ulin, C.R. 2018. *Microsoft's politically correct chatbot is even worse than its racist one*. Quartz. Retrieved October 10, 2022, from https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one.

Sunstein, C. 2004. Democracy and filtering. *Communications of the ACM* 47 (12): 57–59.

Valdez, A., and M. Ziefle. 2019. The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies* 121: 108–121.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems,* 6000–6010. Red Hook, NY: Curran Associates Inc.

Walek, B., and V. Fojtik. 2020. A hybrid recommender system for recommending relevant movies using an expert system. *Expert Systems with Applications* 158: 113452.

Zhang, J. 2011. Anchoring effects of recommender systems. *In Proceedings of the fifth ACM conference on Recommender systems*, 375–378. New York, NY: Association for Computing Machinery.