Interacting with an Inferred World: The Challenge of Machine Learning for *Humane* Computer Interaction

Alan F. Blackwell
University of Cambridge
Computer Laboratory
afb21@cam.ac.uk

ABSTRACT

Classic theories of user interaction have been framed in relation to symbolic models of planning and problem solving, responding in part to the cognitive theories associated with AI research. However, the behavior of modern machine-learning systems is determined by statistical models of the world rather than explicit symbolic descriptions. Users increasingly interact with the world and with others in ways that are mediated by such models. This paper explores the way in which this new generation of technology raises fresh challenges for the critical evaluation of interactive systems. It closes with some proposed measures for the design of inference-based systems that are more open to humane design and use.

Author Keywords
Machine learning; critical theory

ACM Classification Keywords
H.5.2. User interfaces; I.2.0 Artificial intelligence

INTRODUCTION

Anxiety about artificial intelligence (AI) has pervaded Western culture for 50 years - or perhaps 500. As I write, the BBC News headline reads *Microsoft's Bill Gates insists AI is a threat*, while the Cambridge Centre for Existential Risk, led by past Royal Society President and Astronomer Royal Sir Martin Rees, lists AI as the first in its list of the technologies that may represent 'direct, extinction-level threats to our species'. The legend of the Golem and other examples from fiction demonstrate human unease about machines that act for themselves. Of course, this ability is shared by devices as humble as the thermostat, the sat-nav, or predictive text – all somewhat mysterious to their users, and all possessing some kind of 'mind' – that is, an internal model of the world that determines system behavior.

The goal of this paper is to analyze the ways in which our relationship with such devices and systems is changing, in response to the continued developments arising from Moore's Law (namely, the scale and complexity of their internal models) and the economics of network connectivity (the range of data from which models can be derived). One set of concerns about those changes of scale is now familiar as an implication of the phrase 'big data' [9], but I contend that the real problem is more insidious - that it results from changes in the technical underpinnings of artificial intelligence research, which are in turn changing designers' conceptions of the human user. The danger is not the creation of systems that become maliciously intelligent, but of systems that are designed to be *inhumane* through neglect of the individual, social and political consequences of technical decisions.

The structure of the paper is as follows. The first section discusses the nature of the intellectual changes that have accompanied developments in AI technology. For many years, researchers in Human-Computer Interaction (HCI) have been concerned about the consequences of AI, but I argue that it is time for some fundamental changes in the questions asked. The second section takes a specific research project as a focal case study. That project was not designed as an interactive system, but suggests future scenarios of interaction that seem particularly disturbing. I use the case study as a starting point from which several key concerns of humane interaction can be explored, introducing questions of human rights, law and political economics. The final section is more technically oriented, describing practical tests available to the critical technical practitioner, followed by a selection of research themes that might be used to explore more humane interaction modes.

BACKGROUND

In recognition of the long view taken at this conference, I will set out the development of these concerns over a relatively extended time-frame. The phrase big data has only recently become popular as a marketing term, promoting the potential for statistical analysis to model the world by extracting patterns from data that is becoming more readily available. The Machine Learning techniques used to create such models have supplanted earlier AI technology booms such as symbol-manipulation methods in the Expert Systems of the 1980s – now remembered nostalgically as GOFAI, for 'Good Old-Fashioned AI'.

These changing trends are significant to Human-Computer Interaction (HCI) because GOFAI has long had a problematic relationship with HCI – as a kind of quarrelsome sibling. Both fields brought together knowledge from Psychology and Computer Science, to an extent that in the early days of HCI, it was difficult to distinguish HCI from AI or Cognitive Science. The program of work at Xerox PARC that resulted in the seminal book *The Psychology of Human-Computer Interaction* [10] was initiated by a proposal from Allen Newell filed as *Applied Information-Processing Psychology Project Memo Number 1*[25], while Norman and Draper's UCSD book (*User Centered System Design*) [28] punningly emerged from the Cognitive Science department at UCSD (UC San Diego). Many early HCI researchers had taken degrees in AI or Cognitive Science (including the current author). Even today, popular accounts of HCI research, referencing both psychology and computing, often result in the naive assumption that we must be doing AI.

However, the years of the 1980s Expert Systems boom were also associated with a critical reaction. The 'strong AI' position anticipated computers that could pass the Turing Test and simulate people. This possibility was challenged not only by philosophers of mind, but by researchers concerned with HCI such as Winograd and Flores [39], Gill [16], and Suchman [35], all of whom noted the ways in which symbolic problem-solving algorithms neglected issues that were central to human interaction, including social context, physical embodiment, and action in the world. Each of these researchers had distinct concerns, but these can be summarized in their reception by AI researchers as problems of *situated cognition* – the failure of formal computational models of planning and action to deal with the complexity of the real world, where actions seem more often improvised in response to events and situations [1, 29].

Although the phrase 'situated cognition' has become tainted by debate and controversy, these debates between HCI and symbol-processing AI have been underlying concerns of major theoretical contributions in HCI, such as Dourish's discussions of context and embodiment [14, 15], and concepts of interaction offered as design resources in a critical technical practice [2, 20].

The New Critical Landscape

The goal of this paper is to explore the ways in which this continued central tension in HCI is now changing in fundamental ways, because of the technical differences between the methods of GOFAI, and those that are now predominant in Machine Learning (ML). Where symbol-processing approaches failed to take account of the rich information available in the world, ML algorithms have access to huge quantities of such information. This has resulted in enormous changes, by comparison to the technical and commercial context in which the field of HCI was formed.

	1980s-90s	**2000s-10s**
Empirical Motivation	Experimental Psychology	Functional Imaging
Intersectional Community	Cognitive Science	Neuroscience
Technical Research	Artificial Intelligence	Machine Learning
Business Opportunity	Expert Systems	Big Data
Critical Question	Situated Cognition	Humane Interaction

Table 1. Structural analogy of the new technical context for HCI, with the focus of this paper as a new critical question.

To summarize those changes, and also the central analogy leading to the concern of this paper, Table 1 outlines the relationship between the technoscience landscape of GOFAI that laid the context for HCI in the 1980s, and the corresponding scientific, technical and business trends to which we should be responding today.

The critical challenges to symbol-processing GOFAI were that the symbols were not grounded, the cognition was not situated, and there was no interaction with social context. These are not the critical problems of the ML landscape. In contrast to those earlier critiques, machine learning systems operate purely on 'grounded' data, and their 'cognition' is based wholly on information collected from the real world. Furthermore, it appears that ML systems *are* interacting with their social context, for example through the use of big data collected from social networks, personal databases, online business and so on. However, this question of interacting with social data introduces the most important critical concern of this paper.

This concern can be framed in terms of the Turing Test. In that test, judges are asked to distinguish between a human and a computer. In the classic formulation, we wish to see if the computer will appear sufficiently human-like that the two cannot be distinguished. However, I am more concerned with the reverse scenario. What if the human and computer cannot be distinguished because the human has become too much like a computer?

The original version of the Turing Test fails for all the reasons identified in the 1980s HCI critiques of AI. The new version 'passes' the Turing Test, but in a way that demands a new critique. My concern is that reducing humans to acting as data sources is fundamentally inhumane. A serious additional concern is that technical optimists appear to be blind to this problem – perhaps because of their excitement as they finally seem to be approaching the scientific 'goal' of the Turing Test. HCI researchers and other critical practitioners should be alert to these technical developments, and be ready to draw attention to their consequences.

This is particularly important because, whereas Expert Systems attracted much technical excitement in the 1980s, they were not widely deployed – at least, not to the extent that ordinary people would interact with them every day. In contrast, the statistical techniques of ML are now widely deployed in interactive commercial products. Everyday examples include the Pagerank algorithm of Google, Microsoft's Kinect motion-tracking interface, many different mobile predictive text entry systems, Amazon's recommendation engine and so on. If there is a critical problem, it is not simply an academic or philosophical concern about a speculative future of intelligent machines. On the contrary, I believe this may be a fundamental problem for contemporary society.

To summarise this section, it has drawn comparisons between the technical trends that inspired the HCI critiques of the 1980s, and the technical trends of today. Much theory in contemporary HCI originally emerged from those 1980s critiques of symbol-processing AI. But whereas the core problem of symbol-processing AI was its lack of connection to context – the problem of situated cognition – the core problem of machine learning is the way in which it reduces the contextualised human to a machine-like source of interaction data. Rather than cognition that is not *situated*, our new concern should be interaction that is not *humane*.

AN OUTLINE OF THE PROBLEM

The key question is whether ML systems, which create their own implicit internal model of the world through inference rather than an explicit symbolic model, carry new consequences for those interacting with them. There are some existing concerns regarding the epistemological status of such systems.

One is summarised by Breiman [7] as resulting from 'two cultures' of statistical modeling. Breiman contrasts the traditional practice of a statistician building and testing a mathematical model of the world, to ML techniques in which the model is inferred directly from data. He observes that "nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable" [7, p.205]. As a result, rather than following Occam's razor, he believes that "the models that best emulate nature in terms of predictive accuracy are also the most complex and inscrutable" [7, p 209].

Breiman's analysis, if correct (there are objectors), suggests that predictive accuracy is more important than interpretability. Systems built around such models therefore predict the state of the world *and the people in the world, including the actions of the user*, without offering an explicit symbolic explanation. One consequence of collecting data without explaining why is to trigger the now-familiar concerns of surveillance and privacy associated with 'big data'. More subtle is the consequence of interacting with the world through the mediation of a model that purports to explain the user, yet cannot explain itself.

This second concern has been elaborated in debate between Norvig and Chomsky [30]. Chomsky questions the value of models that contain no symbols for us to inspect, while Norvig observes that these models clearly work, and have supplanted symbolic models as a result. As with Breiman's *Two Cultures* essay, this is partly an appeal to technological pragmatism over epistemological reservations – Breiman himself says "the goal is not interpretability, but accurate information" [7, p 210]. However, if this is the case, it is reasonable to ask whether the most *pragmatic* approach will necessarily be the most humane.

These questions are located in a complex nexus of technical and psychological considerations. Before exploring them further, I illustrate that context with a specific case study of research within this nexus. In the following discussion, the case study also serves as a useful (and, by intention, slightly distanced from HCI) concrete example.

CASE STUDY: READING THE MIND

My case study comes from the work of Jack Gallant's research group, described as 'reconstructing visual experiences from brain activity' [27]. Generating wide public interest from the disturbing suggestion that his team had created a mind-reading machine, this ML system retrieved film clips from a database according to the similarity of EEG readings taken while people watched the films. Press reports and publications were accompanied by images such as that in Fig. 1, showing (on the left) a still from the film that had been shown to an experimental subject, together with (on the right) a 'mind-reading' result that was often interpreted as the rendering of an image captured from within the brain of the subject.

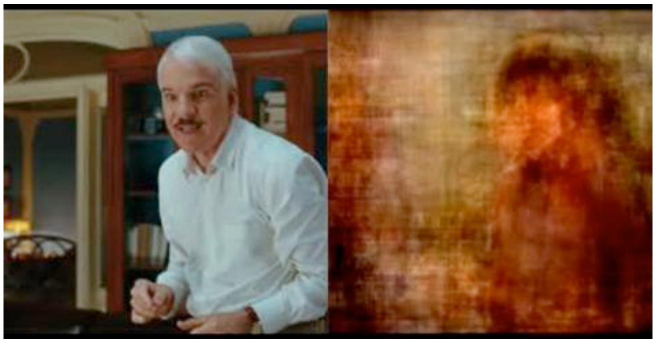

Figure 1. A visual image reconstructed from brain activity, as reported in [27] (image reproduced with permission of the authors, extracted from https://www.youtube.com/watch?v=nsjDnYxJ0bo, and reused under CCA license). The left side of the figure shows a film scene presented to an experimental subject. The right side is a reconstructed image informed by EEG measurements from the subject's brain.

The visual rhetoric in these scientific illustrations is compelling. The captured mind-image on the right of Fig. 1 resembles an oil painting by Goya or by Turner, an artistic rendering of the deep unconscious. The image on the left is a scene from a Steve Martin film – reminding us of Martin's classic riffs on disembodied brains (*The Man with Two Brains*, 1983) and mind-body exchange (*All of Me*, 1984). But on closer inspection, the resemblance between the right and left halves is rather slim. Why does the reconstructed image have dark hair rather than white? Why is it not wearing a white shirt? In fact, all we can say for sure is that there appears to be a human figure on the right of the frame. We are sure because we see its face – as plainly as we see a face in the moon!

Of course, humans (in the Northern hemisphere) observe a face in the moon because human brains, as with the brains of all social mammals, have evolved to detect faces well. If one were to record EEG signals while humans observe a wide range of different images, we might be confident that images of faces would result in some of the most distinctive signals. We also know that each hemisphere of the brain responds to only one half of the visual field, so the simplest possible inference is to determine the side of the head on which a signal has been detected.

So, this is a point at which one might ask more closely how this image of the unconscious was painted. Might it be possible to construct such an image, if the ML model were so crude as to encode only the presence of a face on one side or other of the visual field? It seems, from close reading of the research described, that the right hand image is a blurred average of the 100 film library scenes most closely fitting the observed EEG signal. That is, it is a blurred combination of 100 film scenes in which a face appears in the right hand side of the frame. The other visual features, given this starting point, are unexceptional. The location of the face within the right hand side follows a completely conventional film frame composition. The face is located by the rule of two thirds, and the gaze of the figure is directed inward to the center of the frame. In fact, this could be a scene from any film at all.

The rhetorical skill of the researchers in constructing this striking image is to have chosen just the right number of film scenes to average. A composite of only one or two scenes would retain sufficiently clear visual features to make it obvious if we were not looking at the right film. On the other hand, an average of 10,000 scenes would be no more than a brown smear, losing the powerful hints of contour and shadow that remind us of dream paintings. Even the level of detail used in the averaging process seems carefully chosen – the unit pixels, the hints of vertical lines suggesting scenery – are the size and shape of brush strokes to further evoke painterly rendering.

In this case study, we can admire the skill of the researchers in presenting such a compelling story. However, this example also provides an opportunity to discuss some essential considerations in the critical assessment of ML models. I will not dwell further on the visual rhetoric of scientific discourse, despite the fact that the case raises fascinating questions in that area too.

From Case Study to Critical Questions

This case study has illustrated typical techniques from ML and neuroscience that might (if they worked in the manner implied) provide a basis for new models of user interaction that would be able to predict the user's needs. A 'natural' brain-computer interface of this kind has not yet been proposed – though it may come soon! However, the case study can also be used as a starting point for critical enquiry.

The central part of this paper explores the resulting questions. The following sections draw out these questions in turn, starting with those that arise directly from the domain of film – art works and their readers – but then moving on to the economic and psychological networks in which artworks are embedded.

QUESTION 1: AUTHORSHIP

GOFAI systems maintained a clear distinction between data (symbolic representations of the world), and algorithms that processed this data. In the modern digital economy, business and legal frameworks still maintain a clear distinction between (data-like) content and (algorithm-like) services. However, the *behavior* of ML systems is *derived from* data, through the construction of the statistical model.

The image in Fig. 1 appeared meaningful because it was derived from actual movie scenes. Similarly, ML-based interactive technologies such as the Microsoft Kinect game controller are able to recognize human actions because their models have been derived from a large database of human actions [33]. In one sense, these statistical models can be regarded as an index of the content that created them, allowing the system to look up an interpretation that was originally created by a human author.

This close relationship between index and authorship has been a focus of critical inquiry in the past, for example in Borges' *Library of Babel* [8], which contained every possible book in the universe that could be written in an alphabet of 25 characters. If there were a catalogue to this library, its index would have to contain the full text of the book itself, to distinguish each book from the volume that is identical apart from a single character. Borges' *Library* was a thought experiment, but we do now have ML algorithms that index (nearly) every book in the world, meaning that their models incorporate a significant proportion of those books' content. However, the algorithm collecting an index does not care whether the data is copyrighted – despite the fact that the copyrighted content is in some way mashed-up into the structure of the resulting model.

To consider the implications, imagine a dynamic audio filter trained on a single song – perhaps John Lennon's

Imagine. If allowed to process sufficient random noise, this filter could select enough sounds to reproduce the song. If applied to any other song, it selects only those parts of the song that resemble *Imagine*. The model is not far from being a copy. But what if we trained the model on two Lennon songs, or on the whole Beatles repertoire, to the extent that it selects or simulates that repertoire? This is not dissimilar to the 'reconstructed' film scene described in the experiment above, although we are closer to technical feasibility with audio than with film (e.g. [12]).

The ethics of copyright are a common enough topic in critiques of digital media. However, rather than purely considering the professional arts (loudly defended by copyright holders in the media industries), we should also consider the ways in which every digital citizen is an 'author' of their own identity, because of the ways that persistent traces of our experiences and interactions with others are now extending beyond our own memories into digital representations. The human self is a narrative of experiences and relations with others, and ownership of this narrative is a critical condition of being human. I return to this issue later.

QUESTION 2: ATTRIBUTION

The logic of digital copyright would assert that the content of the original material captured in an ML model or index should still be traced to the authors – the scale of the appropriation is not the key legal point. However, indexing of the Internet is still heavily influenced by rather utopian perspectives derived from champions of the public domain in which it is often asserted that 'information wants to be free' (attributed to Stewart Brand). But if we acknowledge that 'information' represents the work and property of individuals, then 'freedom' might simply mean the freedom to appropriate that work by those wishing to encode it in the form of ML models, especially if there are no copyright holders leaping to defend their license revenue.

Attribution is already problematic in digital media, as a result of postmodern collaging practices – remixes, mashups and so on. Experts in forensic musicology report that court decisions are contingent on the availability of uncontestable symbolic representations [3]. Lyrics are easy to defend. Melodies likewise, so long as they can be written out as notes on a scale. However, reprocessed samples are more ambiguous, and distinctive digital production techniques almost impossible to verify without separate evidence of provenance. The law is a symbolic system, and it works well only with symbolic data.

The commercial logic applied in digital music licensing (in particular, within sample-based genres) is a logic of contamination – the inclusion of any data fragment results in a derived work, meaning that attribution is required and license fees payable. In contrast, the application of processes and algorithms (whether Autotune, or a fuzzbox) does not imply that the inventor of the fuzzbox owns the song recorded with it. If ML models are interpreted (and applied) as processes, this is a challenge for attribution. Although they resemble a purely algorithmic construct, they are also a kind of intertextual content derived from the data used to train them (just as a John Lennon 'filter' represents all possible John Lennon songs, even including those not written, but predicted from his body of work).

This situation is exacerbated by the fact that, when we interact with computer systems, we often over-attribute intelligence to the system, failing to recognize the fact that apparent intelligence, in a system that is in some way faulty or ambiguous, may arise from our own judgments. Collins and Kusch describe this dynamic as 'Repair, Attribution and all That' (RAT) [11]. RAT explains the enthusiastic popular reception of the mind-reading demonstration in Figure 1. Although clearly a poor reproduction of the original scene, the human observer 'repairs' this, interpreting the computer output as an *accurate portrayal of a dream*, and attributing their *own* subsequent interpretation of that ambiguous image as resulting from the apparent intelligence of the *system* that produced it.

In symbolic systems, where the system behavior has been specified and encoded directly by a human designer, the user can apply a semiotic reading in which the user interface acts as the 'designer's deputy' [34]. In contrast, if the system behavior is encoded in a statistical model, derived by inference from the work of many authors, and presented to users in a context where any faults are repaired and attributed to the intelligence of the system itself, then this humane foundation of the semiotic system is undermined.

Models derived from big data defy symbolic representation, because the scale and complexity of the data processing algorithms makes it very difficult (or even impossible) to 'run them backward' and recover the necessary links for human attribution. We can choose to treat such dynamics as a necessary sacrifice for the public good of the creative commons, supporting a massive global project of postmodern collaging. However, if we are accelerating the death of the author through technical means, then who will get paid?

QUESTION 3: REWARD

Attribution of authorship is considered an inalienable human right [37]. However, as noted by Scherzinger [32], the digital public domain pays lip service to attribution of authorship, while actually providing unfettered access to commercial interests. Global indexing and data-centric service models represent a new era of enclosures, echoing the enclosure of common grazing land by the British aristocracy. In particular, Scherzinger notes that there is a tendency for the information assets of the global South to be incorporated into the 'public domain' administered from the North, while the revenue in services derived from those assets continues to flow from the South to the North.

I have already discussed the way in which the separation of data and algorithms in the systems of the GOFAI era has become far less distinct in the statistical models that underlie the behavior of ML devices such as Kinect. There is a commercial analog to this technical change, in the relationship between content and services. In the contemporary digital economy, we retain a notional separation between content and services. However, in practice, the corporations responsible for digital infrastructure 'ecosystems' find it useful to blur those boundaries. Apps for the iPad and iPhone often prevent the user from inspecting stored data other than through the filters of the application itself. The market models of interactive digital products (such as the AppStore) are gradually integrated with those of digitized content delivery (such as iTunes), and the company deploys proprietary services such as cloud storage, user account authentication and so on, on top of these. The ecosystem players – Apple, Google, Facebook and Microsoft – are all attempting to establish their control through a combination of storage, behaviour and authentication services that are starting to rely on indexed models of other people's data.

This is a more serious problem than the commonplace observation that "if you are not paying for it, you're not the customer; you're the product being sold" (e.g. [19]). While national and international legal frameworks focus on outdated models of content protection (through copyright and licensing) and service provision (through free trade and tariff agreements), the primary mechanism of control over users comes through statistical index models that are not currently inspected or regulated. The regulated revenue models of whole industries are being disrupted by digital alternatives that bypass retail distribution with proprietary indexing and access (e.g. Spotify, YouTube and others). The underlying models are transnational, with the corporations increasingly resembling the *zaibatsu* of William Gibson's cyberspace, rendering national jurisdictions irrelevant in comparison with the internal structure of ML models.

This is a concern that is likely to become far more pressing after implementation of the proposed Transatlantic Trade and Investment Partnership, which would allow corporations to sue a country whose laws obstruct their interests. As a result, the significance of this analysis extends beyond the purely commercial, to the political foundations of our economic system [24]. When we replace content with services that obscure individual authorship through the construction of statistical models, Carlo Vercellone [38] observes that we are developing a new form of cognitive capitalism, in which access to the proprietary infrastructure encoded in models and indexes operates as a rent – extracting surplus value from the labor of the majority. The owner of the statistical model used to index and rank content, as in the case of Google's PageRank, thus becomes a rentier, rather than a proprietor, of digital content [31].

QUESTION 4: SELF-DETERMINATION

The previous section has offered a relatively traditional Marxian analysis of what we might consider to be humane, in relation to economic exploitation of many people by a few. However, the original case study also draws attention to the ways in which 'mind-reading' technologies hold implications for the psychological identity of the user. This section considers implications of Machine Learning technologies for the self, as a psychological rather than purely legal and economic construct.

Sense of Agency

The first of these is the perception of agency – the sense of one's self as an undivided and persisting entity, in control of one's own actions. This is a key element of mental health, and is often disrupted in those suffering from delusional syndromes such as schizophrenia. In previous research, we have shown that diagnostic devices used to measure reduced sense of agency in psychiatric illnesses can also be used to measure the user's sense of agency when interacting with ML systems [13].

The behavior of many ML-based systems is determined, not only by models of the external world, but by statistical user models that predict the user's own actions. Those predictions may be based on data collected from other people (as in the case of Kinect), or one's own past habits. However, in all of these cases, one frequent outcome is that the resulting system behavior becomes perversely *more difficult for the user to predict*. Rather than an explicit rule formulated and symbolically expressed by a designer, the behavior is encoded in the parameters of a statistical model – the kind of model that Breiman [7] describes as "complex, mysterious, and, at least, partly unknowable". The result is frequently useful, but can also be surprising, confusing or irritating – as often noted of auto-correct [22].

Whether or not the result is immediately useful, the work of Coyle et al [13] shows that these ML-based predictions reduce the user's sense of agency in his or her own actions. Furthermore, some classes of user may be excluded from opportunities to control the system, because the prior data from which the trained model has been constructed does not take account of their own situation. One such example is those Kinect users whose body shapes are not consistent with the training data of the system, for example because they have absent limbs.

These cases draw attention to the ways in which, when interacting with an inferred model, the user is effectively submitting to a comparison between their own actions and those of other people from which the model has been derived. In many such comparisons, the statistical effect will be a regression toward the mean – the distinctive characteristics of the individual will be adjusted or corrected toward the expectations encoded in the model.

Construction of Identity

This is the second of the 'psychological' problems that I suggest arise when interacting with an inferred world. The processes through which we construct our own individual identity depend on the perception that we are distinct individuals rather than interchangeable members of a homogeneous society. Processes of individuation, in which we separate ourselves from family and from others, are central to the achievement of maturity and personhood. These processes can be damaged through institutional and systemic constraints, leading to a widespread concern for self-determination as a fundamental human right.

If the construction of one's personal identity is achieved substantially through narratives in digital media – through Facebook profiles, photo streams, blogs, Twitter feeds and so on – then the behavior of these systems becomes a key component of self-determination. To some extent, digital media users are highly aware of the need to 'curate their lives' in the presentation of such content. However, they are less able to control the ML models that are inferred from their online behavior. At a trivial level, these models may record unintentional or private actions that the user would prefer to disown. At a more profound level, regressions to the mean result in personal identities that are trivialized (cute animals and saccharine sunsets), or have pandered to a lowest common denominator of mass-market segmentation (prurience and porn).

QUESTION 5: DESIGNING FOR CONTROL

The previous sections have referred to examples of interactive software, although the original case study was not itself proposed as an interactive system. This section considers two specific issues that arise when a user needs to operate products that have been built using ML techniques

Control

The anxieties regarding loss of control in inference-based interfaces are already well-established: small-scale behaviors such as Microsoft's Clippy, Amazon recommendations, or predictive text errors become the object of popular ridicule, tinged only slightly with the anxiety of disempowerment.

However, case studies in which the user's own needs are modeled point to the issues that arise when more complex or larger-scale system behaviors are determined through ML techniques. Since the system behavior is derived from data, if the user wishes to control or modify that behavior, they must do so at one remove, by modifying the data rather than the model. As argued in [4], if the user wishes to change the behavior of the system more permanently, then ML-based techniques can inadvertently make the task *more* challenging. Rather than simply learning the conventions of a scripting or policy language in which to express the desired behavior, the user must second-guess an inference algorithm, trying to select the right combination of data to produce a model corresponding to their needs.

A well-known early example of this challenge was expressed in the *Wall Street Journal* through advice on 'If TiVo thinks you are gay, here's how to set it straight' [40]. The choice of this specific theme drew attention to the implications of surveilling sexual preference, thereby conflating the concerns of privacy with those of control. In later reports, and in a classic meme formulation, the problem has been simplified purely as a matter of surveillance: 'my TiVo thinks I'm gay'. However the TiVo developers at the time experimented with a 'Teach TiVo' function that could be used to modify the inferred model. Although briefly released in a beta version, the company eventually focused on refining the algorithms, rather than offering explicit control to users[1].

Contracts

The problem of how a user can express the behavior they want extends also to the legal relationship between users and service providers. As already noted in the earlier discussion of authorship, the structure of legal frameworks relies on symbolic representation rather than statistical patterns. Service providers now offer end-user license agreements that describe the procedures for collecting data rather than the implications of the model that will be trained from that data. Observation of user behavior, and data-mining from content, have become completely routine, to the extent that it is hardly possible for users to opt out of this functionality if they want the product to work.

At present, these universal license agreements do not help the user to understand what benefits they (or others) will obtain from the resulting inferred models [18]. They also provide no option for customizing or restricting either the model or the contract – the user may opt in or out, but not select any other trade-off between self-determination and convenience. This appears to be a joint failure of technical and legal systems, failing to recognize the interdependence of the two that arises from interacting with the world through an inferred model.

TOWARD HUMANE INTERACTION

The 'mind-reading' case study that provided my initial example has the character of a parlor trick, albeit one that I have used to introduce some genuine ethical and legal issues. To reiterate the real concern for interactive systems: when an ML system sees the world we see, and also observes our responses, the expectation is that it will make predictions about us, and about what we want and need. This inferred model mediates our interaction with the world and with others.

I have drawn attention to numerous ways in which the shift from direct symbolic expression to inferred statistical models of the world has changed the nature of the relationship between interactive digital systems and the

[1] Personal communication, Jennifer Rode, 20 Feb 2015.

people that use them. These include questions about the source of the data in those statistical models, attributions of authorship and agency, the political consequences of shifts to non-symbolic expression, and the effects on the identity of the individual as constructing their own self and controlling their digital lives.

I have suggested that these shifts result in systems that are less humane, because of the ways in which the relationship between the system behavior and human activities has become obscured through the scale and complexity of the modeling process. ML models have become less accountable, open to exploitation by commercial actors, closed to legal inspection, and resistant to the directly expressed desires of the user.

However, the goal of this paper has been to offer a technically-informed critical commentary, as a supplement to the many existing critiques that address more traditional dystopian anxieties of control and surveillance. Rather than simply sounding further warning alarms, or lamenting the loss of a golden age of symbolic transparency, a technically-informed critique should be able to draw attention to opportunities for technical adjustment, caution, correction and allowance.

In the following sections, technical considerations are therefore presented as ways in which the structure of the inferred model might be opened up – more open to understanding by users and to critical assessment by commentators. Where symbolic systems offered direct representations of knowledge, ML systems must be inspected in terms of their statistical structure.

Features

A classic student exercise in the GOFAI days was to ask how a machine vision system might recognize a chair. If the initial answer described four legs, with a seat and a back, the tutor would ask what about a stool, or one with three legs, or a bean bag[2]? Eventually the student might offer a functional description – something a person is sitting on. But then what about a person sitting on a table, or resting on a bicycle? The discussion might end with Wittgensteinian reflections on language, but the key insights are a) that judgments are made in relation to sets of features, and b) that accountability for a judgment is achieved by reference to those features.

In the case of statistically inferred models, the features can often be far more surprising. One of the greatest technical changes in the transition from GOFAI to ML systems has been the discovery that many very small features are often a reliable basis for inferred classification models (e.g. [21]). However, the result is that it becomes difficult to account for decisions in a manner recognizable from human perception. A recent publication illustrated this phenomenon with images such as Fig. 2, and the title 'Deep neural networks are easily fooled: High confidence predictions for unrecognizable images' [26].

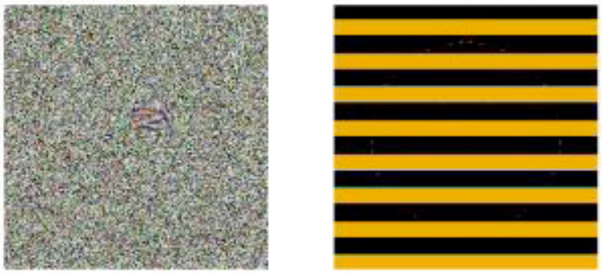

Figure 2. Two images synthesized such that they will result in a classification judgment of 'school bus' (from [26], reproduced with permission of the authors).

Although neither of the images in Fig. 2 is recognizable as a school bus, the image on the right offers a more visually salient account of those features in the training data set which have been encoded in the training process. They allow a human viewer to recall, for example, that school buses in the USA are painted in a characteristic color, and furthermore to speculate that this ML classifier might not be effective in other countries where those colors are not used. In contrast, the image on the left in Fig. 2 does not provide a basis for this kind of assessment (the original paper includes many similar noise-based images that represent different categories, while being indistinguishable to a human viewer).

Contrasts of this kind point toward the design opportunity for more humane ML systems that reveal the nature of the features from which judgments have been derived. It is often the case that such features do not satisfy the symbolic expectations underlying representational user interfaces. The semiotic structure of interaction with inferred worlds can only be well-designed if feature encodings are integrated into that structure.

Labeling

The models underlying many ML-based systems have been constructed from sets of training data, where each example in the data set is labeled according to a so-called 'ground truth,' often an expert judgment (this is characteristic of *supervised learning* algorithms – I discuss *unsupervised* learning below). The inferred model, however complex it might be, is essentially a summary of those expert judgments. However, it is expensive to label large data sets, so the availability of suitable labels often determines the models that can be obtained. For example, in the case of natural language processing systems, the ML model that is most often used to assign a part-of-speech (POS) to individual words (POS-tagging) is based on a training set from the *Wall Street Journal*. When applied to the speech of people who do not talk like the WSJ, the accuracy of this

[2] This example is taken from a class taught by Peter Andreae at Victoria University of Wellington in 1986.

model will be reduced. But it is unlikely that expert judges would invest expensive time labeling (for example) the street patois of a Brazilian favela, even if there were a standard textbook description of its linguistic syntax.

The phrase 'ground truth' implies a degree of objectivity that may or may not be justified for any particular labeled data set. If the interpretation of a case is ambiguous, then that training item must either be excluded from the data set (a common expedient), or labeled in a way that discounts the ambiguity – perhaps because the expert has not even noticed the problem. Furthermore, expert judges may approach a data set with different intentions from those who will interact with the resulting model. One might ask whether these experts are representatives of the same social class as the user, or whether their judgments are dependent on implicit context, perhaps uninformed by situations that they have never experienced.

Moreover, labeling large data sets is tedious and expensive, to a degree that those with broader expertise of judgment might be reluctant to spend their own valuable time in such activity. As a result, many researchers resort to the use of online labor markets such as Amazon's Mechanical Turk (AMT), to commission 'human intelligence tasks'. This strategy casts further doubt on the presumption of ground truth, through the economic relations in which it is embedded. For example, when the AMT service was introduced, it was formally available only to users in North America, with the result that statistical models labeled by AMT workers might incorporate an embedded form of cultural imperialism, perhaps of the kind illustrated in the black and yellow 'school bus' category of Fig. 2.

Confidence

Many of the symbolic models created in the early days of GOFAI were deterministic – a particular output was known to have resulted from a given range of inputs, and those inputs were guaranteed always to produce the same output. In contrast, the behavior of inference-based systems is probabilistic, with a likelihood of error that varies according to the quality of the match between the inferred model and the observed world. This match will always be approximate, because the model is only a summary of the world, not an exact replica. (In fact, training a model until it *does* replicate the observed world is 'over-fitting' – it results in a fragile model that performs poorly at handling new data).

Despite the fact that inferred judgments can carry varying degrees of confidence, many interactive systems obscure this fact. In situations where the behavior of the system results from choosing the most likely of several possibilities, it might benefit the user to know that this behavior resulted from one 51% likelihood being compared to a 49% alternative, as distinct from another case where the model predicts one choice with 99% likelihood. Some of the most successful examples of inference-based interaction, including predictive text and search engines, offer the user a list of choices that have been ranked according to relative confidence. However, these systems do not currently scale the ranked choices in proportion to the magnitude of the prediction. MacKay's *Dasher* is an alternative example of a model-based predictive text entry system that directly exposes the confidence of the prediction in the user interface, by varying the size on screen of the different choices [36].

The challenge for incorporating confidence in an interactive system is to do so unobtrusively, allowing the user to take account of relevant cues without information overload. However, in order to establish this as a design opportunity, we first need to acknowledge that confidence does vary, and that probabilistic inferred models should not be presented as though they were deterministic.

Errors

Decisions made on the basis of an inferred model will include errors. Research results in ML conventionally report the degree of accuracy in the model (80%, 90%, 99% etc). However, the user's experience of such models is often determined by the consequence of the errors, rather than the occasions on which the system acts as expected.

90% accuracy is considered a good result in much ML research, but using such models in an interactive system means that one in ten user actions will involve correcting a mistake. User experience researchers understand the need to focus on *breakdowns*, rather than routine operation, although in the past these have tended to result from indeterminacy in human behavior, rather than in the behavior of the system itself. It is important to recognize that departures from routine are more costly to manage than routine operation, because they require conscious attention from the user [5]. A system that mostly behaves as expected, with occasional departures, may be less useful than one that has no intelligence at all. Furthermore, it is possible that a 1% error rate will be even more dangerous than a 10% error rate, because the operator may become complacent or inattentive to the possibility of error.

Deep Learning

The above discussion of features and labeling applies to the ML research techniques most popular in the early 2000's (and now widely applied in commercial practice), but it should be noted that recent algorithms in the broad category of 'Deep Learning' (including deep belief networks, convolutional neural networks and many others) raise somewhat different issues. Deep Learning techniques aim to be less dependent on explicit feature encoding, and also emphasise the importance of unsupervised learning, so that a labeled training set is not needed. However, each of these attributes leads to further questions for the critical technical practitioner.

The first problem is that, just as it is not possible for a human to gain information about the world unmediated by perception, it is difficult for a Deep Learning algorithm to gain information about the world that is unmediated by features of one kind or another. These 'perceptual' features may result from signal conditioning, selective sampling, optical mechanics, survey design, stroke capture – because every process for capturing and recording data implicitly carries 'features' that have been embedded in the data acquisition architecture through technical means. If the features have not been explicitly encoded as a component of the ML system, then it is necessary for the critic to ask where they *have* been encoded. The questions already asked with regard to obfuscation of the model perhaps become more urgent, in that only the designers of the associated hardware may be able to provide an answer.

The second challenge in assessing Deep Learning systems is that, if the judgments are not made by humans, they must be obtained from some other source. In one of the most impressive applications of convolutional neural networks, the staff of DeepMind Technologies [23] demonstrated a system that can learn to play early Atari video games without any explicit human intervention (other than drawing attention to the score – which is a crucial factor).

Examples of this kind are often discussed with the expectation that the next step after a video game will be action in the real world. Similar assumptions were often made in the GOFAI era, although that focus on 'toy worlds' was eventually abandoned, in recognition that operating 'in the wild' was overwhelmingly more challenging. This quite obviously applies to the case of Atari game worlds, and perhaps such toy applications do not seem a matter for serious concern. However, we do have reason to be concerned if similar algorithms are applied to some of the other representational 'games' played in contemporary society, such as the representational game worlds of corporate finance, audience ratings, or welfare benefits, and the 'scores' that are assigned to them in financial markets or monetarist government.

So critical questions in the analysis of Deep Learning systems can be set alongside those of earlier ML techniques: 1) what is the ontological status of the model world in which the Deep Learning system acquires its competence; 2) what are the technical channels by which data is obtained; and 3) in what ways do each of these differ from the social and embodied perceptions of human observers? Each of these questions represents a deeply humane concern with respect to the representational status of inferred models and the degree to which we are obliged to interact with such models.

Summary

The relationship between inferred models, and the data that they are derived from, is complex. The model is already a summarized version of the original data, although this is not a summary that is directly readable in the manner of a symbolic representation.

It is possible for users to interpret and interact with such models in a way that places more emphasis on human concerns, but this requires designs that communicate essential characteristics of the model. Important aspects include the features that have been used to train the model, the source of the data in which those features were observed, the expert judgments that were applied when labeling the ground truth, the degree of confidence in any particular application of the model, the specific likelihood of errors in the resulting behavior, the infrastructure through which input data was acquired, and the semiotic status of the representational worlds in which an unsupervised model apparently acts.

DESIGN RESOURCES FOR INTERACTIVE ML

This paper has presented an historical argument for a new critical turn in HCI, that steps aside from the preoccupations of symbolic GOFAI and draws attention to the consequences of interacting with the inferred models derived from 'big data'. It has investigated a number of specific problems that arise in such models, where these problems are also attuned to the ways in which distinctions between data and control, or content and services, are changing in the digital economy and regulation of the Internet.

This analysis suggests the need for design considerations that might help users to engage with inferred models in a way that is better informed, more effective, and supports their human rights to act as individuals. We need improved conceptual constructs that can be used to account for a new designed relationship between user intentions and inferred models. The following suggestions are drawn from work carried out in the author's research group, in order to provide concrete illustrations of the kind of design research that might take account of these considerations.

One such construct is the notion of *agency* – if the machine acts on the basis of a world model that is derived from observations of other users (or from the assumptions of an expert labeler), then this will be perceived by the user as a proportionate loss of personal agency through control of one's own actions. Fundamental human rights of identity, self-determination and attribution are implicated in this construct. If the inferred model obfuscates such relations, then they should be restored through another channel.

A second construct is the interaction style previously described as *programming by example* – where future automated behaviors are specified by inference from observations of user action. Often promoted as an idealised personal servant, many such systems struggle to allow the basic courtesies of personal service, such as asking for confirmation of a command, or responding to a change of mind. Empowering users through such techniques will involve explicit representation of the inferred requirements and actions.

A third construct is the recognition that, although approximate and errorful inferred models of the user's intentions are problematic and worrisome, humans themselves also develop world models on the basis of incomplete and selective data. Kahneman's investigations of heuristics and biases in human decision-making [17] offer a mirror to the inferred world model of ML systems. There is a valuable opportunity to create user interfaces that acknowledge and support such *human reasoning styles*, rather than attempting to correct the user on the basis of unseen data or expert design abstractions.

A fourth construct is to reconsider *the role of the state*, in an era when neither intellectual property nor legal policies need be explicitly formulated as symbolic representations. The new status of content that underlies inferred models throws new light on the role of public service broadcasters, who should be in a position to establish and protect genuine public value in the public domain [6].

These four illustrative examples are not proposed as the basis for a unified theoretical framework to be adopted by future design researchers. The intention is rather to provide a relatively pragmatic set of observations and suggestions, showing connections between the ideas in this paper and established topics within mainstream interaction design and digital media studies. Hopefully there are many other such opportunities, which may indeed come together to offer a basis for future design frameworks and methods.

CONCLUSION

The central technical assumptions that underpinned the design of software applications for the first 50 years of the computer industry are now largely outdated. The intellectual agenda of data processing and communications, in which users either interact with each other or make choices between defined system functions, has not been succeeded by the autonomous human-like AI that was anticipated in the 1950s. Of course, HCI has always resisted such ambitions, drawing attention to the pragmatic human needs of social conversation and embodied usability.

In the new technical environment of the 21st century, users increasingly interact with statistical models of the world that have been inferred from a wide variety of data sources rather than explicit design judgments. This situation forces us to attend to the politics of information and action, as well as the attributes and limitations of the inference systems themselves. Just as the technical competence required of engineers is shifting from data and algorithms to information theory and stability analysis, so user experience designers must reconceive the relationship between content and services as constituting an 'inferred world' that stands in rich semiotic relation to individual and collective experience. Doing so requires a philosophical framework in which labour, identity and human rights are recognized as central concerns of the digital era – concerns that are directly challenged by recent developments in engineering thinking. In short, we need a discipline of humane computer interaction.

ACKNOWLEDGMENTS

I am grateful for feedback and suggestions from Luke Church, Moira Faul, Leonardo Impett, Elisabeth Leedham-Green, David MacKay, Becky Roach, Advait Sarkar, and anonymous reviewers of earlier versions.

REFERENCES

1. Agre, P. (1997) *Computation and Human Experience*, Cambridge University Press.
2. Agre, P. E. (1997). Towards a critical technical practice: lessons learned in trying to reform AI. In Bowker, G. Star, S. L & Turner, W. (Eds) *Social Science, Technical Systems and Cooperative Work*, Lawrence Erlbaum, Mahwah, NJ, pp. 131-157.
3. Bennett, J. (2013) Forensic musicology: approaches and challenges. In *The Timeline Never Lies: Audio Engineers Aiding Forensic Investigators in Cases of Suspected Music Piracy*. Presented at the International Audio Engineering Society Convention. New York, USA, October 2013.
4. Blackwell, A.F. (2001). SWYN: A visual representation for regular expressions. In H. Lieberman (Ed.), *Your wish is my command: Giving users the power to instruct their software*. Morgan Kauffman , pp. 245-270.
5. Blackwell, A.F. (2002). First steps in programming: A rationale for attention investment models. In *Human Centric Computing Languages and Environments, 2002. Proceedings. IEEE 2002 Symposia on* (pp. 2-10). IEEE.
6. Blackwell, A.F. and Postgate, M. (2006). Programming culture in the 2nd-generation attention economy. Presentation at *CHI Workshop on Entertainment media at home - looking at the social aspects*.
7. Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science* 16(3), 199–231
8. Borges, J.L. (1941/tr. 1962) The Library of Babel. Trans. by J.E. Irby in *Labyrinths*. Penguin, pp. 78-86
9. boyd, d. and Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, **15**(5), 662-679,
10. Card, S.K. Allen Newell, and Thomas P. Moran. 1983. *The Psychology of Human-Computer Interaction*. L. Erlbaum Assoc. Inc., Hillsdale, NJ, USA.
11. Collins, H. and Kusch, M. (1998). *The Shape of Actions: What Humans and Machines Can Do*. MIT Press.
12. Cope, D. (2003). *Virtual Bach: Experiments in musical intelligence*. Centaur Records.

13. Coyle, D., Moore, J., Kristensson, P.O., Fletcher, P. & Blackwell, A.F. (2012). I did that! Measuring users' experience of agency in their own actions. *Proceedings of CHI 2012*, pp. 2025-2034.

14. Dourish, P. (2001). *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press.

15. Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing* **8**(1), 19-30.

16. Gill, K.S. (Ed.) (1986). *Artificial Intelligence for Society*. Wiley.

17. Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

18. Gomer, R., schraefel, m.c. and Gerding, E. (2014). Consenting agents: semi-autonomous interactions for ubiquitous consent. In *Proc. Int. Joint Conf. on Pervasive and Ubiquitous Computing:* (UbiComp 14).

19. Kepes, B. (2013). Google users - you're the product, not the customer. *Forbes Magazine*, 4 December 2013.

20. Leahu, L., Sengers, P., and Mateas, M. (2008). Interactionist AI and the promise of ubicomp, or, how to put your box in the world without putting the world in your box. In *Proc. 10th int. conf. on Ubiquitous computing (UbiComp '08),* pp. 134-143.

21. Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc 7th IEEE Int. Conf. on Computer Vision*, pp. 1150-1157.

22. Madison, J. (2011). *Damn You, Autocorrect!* Virgin Books.

23. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). *Playing Atari with deep reinforcement learning*. http://arxiv.org/abs/1312.5602

24. Monbiot, G. (2013). Transatlantic trade and investment partnership as a global ban on left-wing politics. *The Guardian*, 4 Nov 2013. http://www.monbiot.com/2013/11/04/a-global-ban-on-left-wing-politics/

25. Newell, A. (1974). Notes on a proposal for a psychological research unit. *Xerox Palo Alto Research Center Applied Information-processing Psychology Project. AIP Memo 1*

26. Nguyen, A., Yosinski, J. and Clune, J. (2014). *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*. arXiv:1412.1897 [cs.CV] http://arxiv.org/abs/1412.1897

27. Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641-1646.

28. Norman, D. A., & Draper, S. W. (1986). *User centered system design*. *Hillsdale, NJ.*

29. Norman, D.A. Cognition in the head and in the world: an introduction to the special issue on situated action. *Cognitive Science* 17, 1-6 (1993).

30. Norvig, P. (2011). *On Chomsky and the two cultures of statistical learning*. http://norvig.com/chomsky.html

31. Pasquinelli, M. (2009). Google's PageRank algorithm: a diagram of cognitive capitalism and the rentier of the common intellect. In K. Becker & F. Stalder (eds), *Deep Search*. London: Transaction Publishers.

32. Scherzinger, M. (2014). Musical property: Widening or withering? *Journal of Popular Music Studies* 26(1), 162-192.

33. Shotton, J., T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM* 56(1), 116-124.

34. de Souza, C.S. (2005) *The Semiotic Engineering of Human-Computer Interaction*. The MIT Press.

35. Suchman, Lucy (1987). *Plans and Situated Actions: The Problem of Human-machine Communication*. Cambridge: Cambridge University Press.

36. Ward, D.J., Blackwell, A.F. & MacKay, D.J.C. (2000). Dasher - a data entry interface using continuous gestures and language models. In *Proc. UIST 2000*, pp. 129-137.

37. United Nations General Assembly. (1948). *Universal Declaration Of Human Rights*, Article 27.

38. Vercellone, C. (2008). The new articulation of wages, rent and profit in cognitive capitalism. Paper presented at *The Art of Rent* Feb 2008, Queen Mary University School of Business and Management, London.

39. Winograd, T. and Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Intellect Books.

40. Zaslow, J. (2002) If TiVo thinks you are gay, here's how to set it straight. *Wall Street Journal* online Nov. 26. http://www.wsj.com/articles/SB1038261936872356908