Shortcuts to Artificial Intelligence

Nello Cristianini

October 2019

Abstract

The current paradigm of Artificial Intelligence emerged as the result of a series of cultural innovations, some technical and some social. Among them are apparently small design decisions, that led to a subtle reframing of the field's original goals, and are by now accepted as standard. They correspond to technical shortcuts, aimed at bypassing problems that were otherwise too complicated or too expensive to solve, while still delivering a viable version of AI. Far from being a series of separate problems, recent cases of unexpected effects of AI are the consequences of those very choices that enabled the field to succeed, and this is why it will be difficult to solve them. In this chapter we review three of these choices, investigating their connection to some of today's challenges in AI, including those relative to bias, value alignment, privacy and explainability. We introduce the notion of "ethical debt" to describe the necessity to undertake expensive rework in the future in order to address ethical problems created by a technical system.

1 Introduction

Science and especially technology are partly shaped by social and cultural elements, including practices that are so commonly accepted that are not noticed or questioned. These are often conveyed through exemplar stories of good practice in a field, which Thomas Kuhn has called "paradigms". Paradigm shifts are remarkable moments of scientific creativity, but also have consequences beyond science. We review three crucial 'mutations' that are behind the current paradigm of Artificial Intelligence (AI), and we argue that they are also responsible for the current 'ethical debt' of the field, a concept that we introduce to indicate the necessity to undertake expensive reworking in order to address ethical concerns.

A series of concerns have emerged over the past few years, following the widespread deployment of services powered by Artificial Intelligence. These include ubiquitous - and often invisible - software agents that make personalised decisions, from the recommendation of news items or videos to the filtering of emails, among others. These concerns are often treated as design flaws that can be separately addressed, and research has already begun in that direction. We

argue that they are instead the direct result of the paradigm shift undertaken by AI two decades ago, more specifically of three important shortcuts that enabled the present methodology to develop. We term the cost of reworking the systems into a state that is compliant with current social expectations "ethical debt" - in analogy with the established notion of technical debt in software engineering - and we trace its cultural origins to the the very paradigm shift that led to the present version of AI, which we call "data-driven". To explain this, we need to briefly recapitulate the recent stages of AI evolution, and observe how "technology" does not refer just to an algorithm, but rather to the complex of people, norms, algorithms, data and infrastructure that are required for any of these services to exist. Addressing the current challenges in AI may require adapting all of the above.

The first truly viable and profitable form of Artificial Intelligence has gradually emerged over the past two decades, but it has been just over the past few years that its social impact has been felt, as a result of its pervasive deployment. Both a unified data infrastructure and various AI technologies had to co-evolve, before they could really benefit from each other, and then benefit society. As AI found its place in our lives, we have become more aware of problems with this technology: mass surveillance, personalised targeting of adverts, disinformation, biased decision making and unexplainable decisions.

In order to understand where those 'pathologies' of AI originated from, we need to review how AI came to take its present shape, from the dream of a reasoning computer to the reality of statistical data-driven systems based on the web. As we briefly recall this journey we can point out some crucial shortcuts that allowed the field to move faster - and more cheaply - towards viable and profitable products. These shortcuts were part of an important paradigm shift that the research community underwent, about 20 years ago. The recent story of AI is the story of how we avoided building expensive models of phenomena which we do not yet understand, such as language and vision, contenting ourselves with just emulating specific 'skills' (such as spell-checking or handwriting recognition) by exploiting statistical correlations found in large masses of data. Machine-learning algorithms and large masses of data could be used to find those valuable patterns.

This shifted the focus of researchers away from modeling the behaviour or skill to be implemented (perhaps by understanding its underlying mechanisms), and towards securing vast amounts of observations of that behaviour, which could be used as training data for statistical learning algorithms.

This new problem - of collecting training data - was in turn bypassed by the practice of using data sourced "from the wild", an expression that indicates data which was already pre-existing as a byproduct of other activities. But the problem remained - in many cases - of annotating this data, in order to inform the agent about its intended behaviour: autonomous agents only have goals because of some definition of success or failure, which is either provided by the designer, or acquired directly from observing the user in the case of learning agents. While this calls for some form of user feedback, this need was bypassed too, by making use of various proxies (eg using click-through rates as a proxy

for user satisfaction), generally called 'implicit feedback'.

Taken all together, these and other shortcuts enabled us to generate a version of autonomous agents at a very low immediate cost. We now have to face the longer term cost of those decisions, which caused part of the "ethical debt" built into our AI infrastructure. Ensuring the fairness of machine decisions, their transparency, the privacy of users, compliance with new regulations, and securing services against surveillance or hostile manipulations, all will come at the significant cost of reworking the technology at a fundamental level. And in some cases it is conceivable that we might be unable to provide equivalent services in a socially acceptable way - in this case the trade offs between accuracy and social constraints will need to be clearly communicated to lawmakers and the public, so that decisions can be made in the appropriate venues.

In this chapter we review the key design decisions that led to the current version of AI, introduce the notions of 'ethical debt' and 'data supply chain' as key concepts to develop safer AI, propose a distinction between explicit and implicit information used to analyse AI systems, as well as the notion of 'misaligned proxies'. We hope that identifying and naming problems and practices is helpful in the process of critiquing and regulating this important area of research. In no way is this review intended to criticize the scientific contributions discussed in this chapter, just to offer a critique with the benefit of hindsight.

2 Shortcuts and Debt

Current AI exists in symbiosis with the unified data infrastructure that has emerged over the past two decades, combining and replacing previous infrastructures for telecommunications, banking, retail, and more. As always in technology studies, we cannot separate technical innovations from their social context. Business models for AI and Web co-evolved and resulted in very profitable systems powered by autonomous agents, which interact on a daily basis with millions of people. That is where the first problems started emerging, so we will examine the cultural steps that took us to this moment.

While the notion of 'paradigm shift' is well known, and we have already discussed it in this context [9, 8], we should introduce a lesser known concept that will be useful in our discussion.

Technical Debt is a notion used in software engineering to describe the additional cost that will have to be paid in the future as the result of taking a shortcut when developing a software system. In was introduced in 1992 [15] by Ward Cunningham, who noticed the many analogies with an actual debt: while this could be part of a deliberate strategy, in situations where fast deployment is desired, it can also accumulate interest, making it increasingly difficult to implement changes later on. Taking shortcuts essentially borrows from the future, when essential rework will be needed.

A related concept in economics is that of Externality, referring to the cost incurred by a third party which has no control on the creation of that cost [4]. These costs can be either individual or of an entire society, and emerge as the

result of them consuming a service.

We use these terms to observe that AI is in a state of "ethical debt", which we define as a technical debt where the future costs are not due to technical sustainability issues, but to the need to address ethical issues such as externalities imposed on the users. This article attempts to identify the origins of AI's ethical debt, which are both technical and cultural, so that addressing them will require more than just technical solutions. This is the story of how we tried to have a free lunch and ended up in debt, - and of how we might go about paying it back.

3 Stages of a Paradigm Shift

The way AI was imagined and pursued by its pioneers involved discovering some grand principles - say, for example, those underlying natural vision or language - and then using them to design artificial systems to complete some tasks related to vision or language. This approach was based on how other forms of engineering worked: for example space engineering, where the design follows an explicit understanding of the underlying physical laws. Identifying and using the mechanisms behind a skill or behaviour implied assuming a form of causal thinking; for example machine translation should follow after having understood the mechanisms behind the generation of language.

Much of early AI research focused on variations of logical reasoning, taking reasoning to be a fundamental part of generating intelligent behaviour, and using the chess-match as a powerful metaphor for how reasoning should work: cleverly examining alternative courses of action and selecting the most promising one. The necessary domain knowledge could be provided to them, possibly as explicit rules and axioms [11, 10, 8].

That approach to AI, which lasted for decades, failed to deliver viable translation or vision systems, and currently is not what powers the recent success stories of AI. Instead, starting from the late 1990s, researchers in an increasing number of domains settled for a practical approach that allowed them to bypass the frustrating attempt to list explicit rules behind complex phenomena such as vision or language.

By collecting large amounts of training data, they were able to use statistical learning algorithms to produce the desired behaviour, or a version of it. For example, in this way they could develop systems to recognise images of handwritten digits, to translate text, and to flag spam emails.

This alternative approach produced its own set of success stories starting from machine translation [3] and handwriting recognition [28], and then spreading to a number of tasks, including product recommendation, spelling correction, spam filtering. Every time, researchers discovered that it was possible to complete a task without the need for the computer to explicitly represent, or understand, the contents of text or images, nor any underlying mechanisms, by just exploiting statistical correlations found in the training data, to obtain a sort of implicit model. This implicit model was formed mostly by large amounts

of data, and by simple statistical rules that allow this data to be exploited. [1]

What resulted was a new generation of data-driven AI systems, which did away with explicit representations, complex algorithms, and all previous expectations, and just focused on the generation of the required behaviour in machines, by exploiting subtle statistical correlations in vast amounts of data. Machine learning, optimization and data took centre stage, as described in the book [29], and a rush began to gather valuable data [11, 10].

Changing the definition of what counts as 'success' in science is the very essence of a paradigm shift, and this definition is often implicitly encoded in the success stories celebrated in a community or taught to students. We have already noted how Artificial Intelligence underwent a paradigm shift about two decades ago [9], by giving itself permission to take certain shortcuts and therefore - implicitly - to subtly redefine its goals.

Some of the lessons, and a beautiful account of that mindset, can be found in [19], as well as in [29, 8]. Importantly, that new set of success stories also suggested a series of shortcuts to be followed when designing intelligent systems.

Correlation vs. Causation. One important consequence of training statistical algorithms to emulate the decisions or behaviours of humans (eg recommending a book) is that we no longer value the reason why the decision is made, so long as the action it generates is appropriate. Predictions count more than explanations, knowing 'what' counts more than knowing 'why', and - as summarised in [29] - 'correlation trumps causation'.

By this slogan, the authors meant that (in the practice of AI as well as other areas of science) there was a change: a focus on establishing and exploiting causal links is replaced by a focus on establishing and exploiting correlational links. While this had been traditionally seen as a fallacy, it became common practice in those domains where the traditional (hypothetical-deductive) method had not worked.

This position was best described in [1], a popular and influential article which hailed "the end of theory", which was brought about by applying data-driven methods to science. That article summarised this shift as: "*(...) faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete (...) There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. " The article supports its provocative claims by noting that - in order for Google to recommend pages - "no semantic or causal analysis is required".*

One year later, the paper [19] notes: *"early work on machine translation relied on elaborate rules for the relationships between syntactic and semantic patterns in the source and target languages. Currently, statistical translation models consist mostly of large memorized phrase tables that give candidate map-*

[1] This step also relates to statistical use of "non-parametric" models in lieu of parametric ones: nonparametric models differ from parametric ones in that the number and nature of the parameters is not fixed before seeing the data. Most modern machine-learning methods are nonparametric, and therefore they do not use the data as in classical statistics to estimate the value of a variable with a precise meaning within a theoretical framework.

pings between specific source- and target-language phrases. Instead of assuming that general patterns are more effective than memorizing specific phrases, today's translation models introduce general rules only when they improve translation over just memorizing particular phrases (for instance, in rules for dates and numbers)."

While this shortcut saves the enormous cost of understanding and explicit modelling, it creates another cost, which is that of sourcing vast masses of relevant training data, and there is no reason - a priori - to expect that this cost should be any smaller. Generating, curating and annotating high quality data is a significant expense in several industries, eg in drug testing. This cost was also bypassed by the AI industry.

Data from the Wild. The second shortcut was memorably summarised in the paper [19] which draws general lessons from the success stories of speech recognition and machine translation. It identifies the causes for those successes in the availability of large amounts of data, already created for different purposes. *"In other words, a large training set of the input-output behaviour that we seek to automate is available to us in the wild. In contrast, traditional NLP problems such as (. . .) POS tagging (...) are not routine tasks so they have no large corpus available in the wild. Instead a corpus for these tasks requires skilled human annotation. Such annotation is not only slow and expensive to acquire, but also difficult for experts to agree on (...). The first lesson of web-scale learning is to use available data rather than hoping for annotated data which is not available. For example we find that useful semantic relationships can be learned from the statistics of web queries, or from the accumulated evidence of web-based text patterns and formatted tables, in both cases without needing any manually annotated data"*

Data gathered from the wild has been crucial in the design of object recognition systems [16], face recognition [20, 21], machine translation [24], etc. The ubiquitous word embeddings that allow us to represent the meaning of words before we process them, are also all learned from data gathered from the wild [30].

Having replaced modeling with data, and replaced generating data with collecting it from the wild, takes AI designers very close to a free lunch, but not quite all the way there. Often a learning algorithm needs to be told what to do, and this comes in the form of supervision, or feedback. The user should tell the agent which of its decisions was appropriate, and which was not: this is a form of data annotation, or curation, that communicates to the agent its intended behaviour.

Proxies and Implicit feedback. A further cultural step addressed this problem. Rather than asking users explicitly what they wanted the AI system to do - a chore that many users are reluctant to take on - designers started making use of implicit feedback, which is another way to say that they replaced unobservable quantities with cheaper proxies. This started early, for example the paper [2] explains how to design document retrieval systems: *"we make a design decision not to require users to give explicit feedback on which hits were good and which were bad (. . .) instead we simply record which hits people*

follow, (. . .) because the user gets to see a detailed abstract of each hit, we believe that the hits clicked by each user are highly likely to be relevant (. . .)".

It is reasonable to expect this proxy to be somewhat aligned with the elusive quantity of "relevance to the user", but understanding the misalignment between a proxy and the intended target has become an important question for AI. This happens particularly as 'retrieval' is replaced by 'recommendation', and the business model of user 'satisfaction' is replaced by that of user 'persuasion'. Is the goal of the agent to maximise relevance, or just click-through rates? [2]

The shift between retrieval and recommendation is very subtle, as they rely on the very same set of techniques. After being used in retrieval, implicit feedback was also proposed as a way to improve recommender systems since 1998 [31], and clickthrough data were proposed since 2002 as a proxy for relevance in search engines [23]. From the late 1990s Amazon and others were making use of the feature "people who bought this also bought . . . " which also makes a clever use of implicit signals [34].

In each of these cases, the assumption is that the user's actions reveal their preferences, or needs, as well as (or even better than) would be done by an explicit feedback. A problem that we need to address is the consequence of using misaligned proxies in training autonomous agents.

Samples of user behaviour were first used by agents to learn general phenomena, such as correct spelling. Then they were used to link the most relevant hits to a given query. Finally they were used to infer an individual's user preferences. Along the way, incidentally, the focus started shifting from serving the users to serving the advertisers. With each of those small decisions, the framing of the goals in AI was slightly shifted, from causation to correlation, from retrieval to recommendation, from understanding to behaving, from serving users to steering them, from making the data you need, to adapting your needs to whatever data you can find.

These steps are what enabled a very low cost of entry for anyone to develop AI agents: the statistical algorithms were simple, the data was available in the wild, the curation or feedback signals were provided by the users implicitly. Though each of these steps introduced assumptions and approximations, these shortcuts promised us a nearly free passage to Artificial Intelligence: by the 2010s we were able to train systems to recognise (or guess) faces, intentions, speech, recommendable products, unwanted emails, interesting videos, all based on fully automated analysis of behavioural or other personal data that was freely available online.

By that time we also had a global infrastructure through which these agents could gather the data they needed, and they could provide very valuable services. The users were presented with a new mass medium, that is constantly looking back at them trying to guess their intentions, that learns from their behaviour, and that remembers everything.

What came next was a lesson in externalities, unintended consequences and

[2] Although this definition is informal, it is possible to create formal definitions of alignment between two learning tasks, eg [14].

technical debt. Part of the problem was that we built AI agents by allowing them to learn from data collected from the wild, and annotated by observing human behaviour. But the other part of the problem was the special place that those AI agents came to occupy within our data ecosystem. Before we discuss the side effects of those shortcuts, we will briefly turn our attention to the data ecosystem within which AI was deployed, as that is also part of the "recipe" for current versions of AI.

4 A New Medium

The emergence of data-driven AI is directly connected to the emergence of a unified data infrastructure, which is larger than just the World Wide Web, but which we will occasionally call - for convenience - just Web (sensu lato). This infrastructure resulted from the rapid convergence of various elements, and the introduction of new ones: computer and telephone networks, personal computers, mobile phones, ATM networks, and the various layers of hardware and software that underpin all that, from the oceanic cables on the one end to the personalised recommendation software on the other end. Importantly, an understanding of this infrastructure must include cultural components too, such as the legal systems regulating its parts, the business models, and social acceptance.

For example: when we buy a book from Amazon, we may have accepted the personalised suggestion of its recommender software, then we complete an online payment, at which point on the other side a combination of robots and people dispatches the book from a warehouse. International tax regulations, labour laws, publishing laws, telecommunications and banking, all of these components must be in place, before one can imagine such a service.

The rapid convergence of this global infrastructure created an ecosystem for autonomous software agents to thrive: within that context they can have the data they need to learn and improve, the necessary computing and physical infrastructures, and the affordances that allow them to have a viable business model.

It is within the context of the Web (sensu lato) that Intelligent Agents first left the laboratory and ventured into the wild, finding a place within society and becoming part of our daily lives. From there, they started to spread beyond. But by then, Intelligent Agents looked nothing like we had imagined. Their behaviour was generated by statistical signals discovered in vast masses of data sourced from the wild, for which they were constantly hungry, and by cleverly including humans as essential parts of their function, to annotate this data with implicit signals [12].

The recommender agent that we find within Amazon (or Facebook, or a spam filter) includes not only a statistical algorithm and a vast database of past transactions, but also information about transactions that did not take place: items that were suggested but not selected by users, time of day and other contextual information, and so on. Truly it depends on samples of human

behaviour to learn either what we want it to do (as in the spam filter) or what it can make us do (as in the shopping recommendations). The difference between the two is often just in the eye of the beholder: revealing preferences is not different than revealing weaknesses [6] and the agent has no way of distinguishing between the two.

So current intelligent agents include within themselves much more than an algorithm: they include datasets of choices we made, and people who can be used to test future conjectures, in this way providing crucial annotation for future items, which are themselves often produced by humans too. Recommending videos, news items, books, blocking emails, correcting spelling, are all common examples of this symbiosis.

By the 2010s the convergence between data-driven AI and the Web was so tight that neither could have existed without the other. They had co-evolved to a point of symbiosis.

We would be wrong to think of this new mass medium as a sort of modern telegraph: far from being a passive communication line, it looks back at us, partly understands the content of our communications and guesses our intentions. Importantly, humans are both users and also participants of these systems [12, 13].

5 The Recipe

The "secret sauce" that powers the current version of AI has an essential ingredient: samples of human behaviour, often in the form of microchoices performed by millions of users, to be used as proxies for more expensive signals; other ingredients include statistical learning algorithms; a powerful infrastructure for the collection of data and the delivery of services. Statistical learning algorithms detect valuable patterns in the myriad signals generated by users' behaviour, these patterns are used to shape the actions of the macroscopic system, and the infrastructure is used to deliver services, generating value as well as further interaction.

We would be missing an important point if we imagined that the seat of intelligence here is the algorithm, in fact algorithms are often changed within current AI systems, without the users noticing. The intelligence is at the level of the overall system, which is also robust to changes in its participants, and its contents. For example, a video recommendation system owes its behaviour to its current users, its historical data, its current contents and a (set of) machine learning algorithm. Each of them can change in time, without the agent necessarily changing,

6 Consequences: a Rude Awakening

The recipe that gave us this version of AI involves replacing causal links with correlations, explicit models with statistical correlations, cured training exam-

ples with data from the wild, and explicit data annotations with implicit signals and other proxies.

It rests on the strong assumption that our actions reveal information, such as our preferences or factual knowledge of the world. It does not seem to consider the possibility that our actions might also reveal our weaknesses and biases. Yet the field of Computational Social Science has known for a long time that data about online human behaviour does contain them in abundance [35, 27, 17, 22]. When an algorithm is mining terabytes of text, to "triangulate" the meaning of a word such as 'nurse' or 'pilot' from the statistics of its everyday use, how can it avoid also picking up subtle biases in the way people perceive those jobs, e.g., in terms of gender? When another algorithm is observing a user interacting with a video recommendation website, can it really distinguish the information needs of the user from their weaknesses, i.e. things that they do not reflectively endorse, but cannot resist clicking on? [5, 6].

Considering this, it is not that surprising that unintended effects were observed, when the first such agents were deployed at the centre of the global data infrastructure, and expected to recommend news, target ads, screen applications, filter emails, and generally make meaningful decisions. One class of problems followed directly from the need that modern AI has to keep track of our online behaviour, to generate valuable data, while other problems emerged from the use it made of this data: to learn word representations, user preferences, or how to make decisions.

Privacy. In June 2013 The Guardian and the Washington Post reported that intelligence agencies have access to the data of US internet companies under a surveillance programme called Prism. This includes emails, live chats and search histories. The revelation created a scandal, attracting attention to the mass of personal data those companies stored. In 2017 the Observer and the New York Times revealed that a Cambridge Analytica had used user information from Facebook in order to extract psychometric information and better target electoral ads. This revealed to the public the possibility that personal data can be used both to manipulate behaviour and to infer sensitive private information [5]. Insurance companies attempted to launch products whose price depends on the personality of users, as inferred from their social media profiles. The public became aware that over the years, an industry based on data-brokers has emerged for the trading of valuable data about online behaviour.

Bias. Decisions made by algorithms have been suspected to suffer from biases [33]. For example, in 2017 it was reported that job ads targeted at women had a lower pay rate than jobs targeted at men; and that ads targeted at ethnic minorities in the US contained racial biases. In 2017 it was also observed that the word embeddings inferred from text collected in the wild contained evidence both of gender and racial bias, a finding that might help explain the above reports, when considered alongside documented topic-differences in gender bias in the media [17, 22, 27]. All this would be picked up by a statistical algorithm trying to learn the meaning of words [7, 35].

Manipulation. Recommender systems, such as those who propose videos or news, are typically designed to try to maximise some measure of engagement,

often click through rates. This is part of the implicit feedback shortcut. But concerns are emerging that these systems might generate filter bubbles [32] or induce excessive use of media. The use of proxies does not allow the agents to distinguish why a user engages with an item: are users' clicks revealing their preferences, or are they revealing their weaknesses? In the second case, we would have built a learning agent specialised in detecting and exploiting the individual weaknesses of its users [5]. This is related to the problem of value alignment: what the user wants, and what relevant annotation the machine can actually find, are different quantities [18]. It is possible that designing recommenders that only rely on explicit and direct communication from users would solve this problem. Of course this would then suffer from the problem that some users do not want to send this information to the agent, but possibly this should be an option available to the users. Problems can also result from machines accessing psychometric information at individual level, rather than at collective level [26].

Transparency. As we replaced explicit modeling, and representation of causal links with predictions based on implicit statistical correlations, it is very difficult for users to understand the reasons behind the decisions of a machine. It is actually also very difficult to imagine how a machine can give such an explanation, yet this has now become a legal right of European users. The first shortcut is at work here, along with the other two - since the real reasons for a machine's decision depend both on the algorithm and on the data. Accuracy of predictions has been privileged over other aspects of knowledge, in this way subtly redefining what knowledge and models are for. Explainable models might be less accurate, but more acceptable in specific domains. Either way, a trade-off is likely to emerge.

7 Conclusions

What has been accomplished by the research community of AI over the past 20 years is remarkable, and it is not our purpose to point fingers or criticise individual contributions to the field. With the benefit of hindsight we can however reflect on how we introduced assumptions in our systems that are now generating problems, so that we can work on repairing and regulating the current version of AI. The same methods and principles can be perfectly innocuous in certain domain, and become problematic only after being deployed in different domains. This is the space where we will need better informed regulation.

Science and technology are partly shaped by social and cultural elements, including practices that can be so commonly accepted as to go unnoticed. These are often conveyed through success stories, which exemplify what and how a scientific community should be doing. Thomas Kuhn called those stories 'paradigms' defining them as "universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of practitioners" [25].)

Towards the turn of the century, the field of Artificial Intelligence has undergone a 'paradigm shift', changing the exemplar success stories that define

its quest [9, 8]. The current paradigm of Artificial Intelligence has been variously described as focused on accurate predictions rather than accurate models; on detecting correlations rather than causation; on knowing 'what' rather than knowing 'why'. The use of so-called "non parametric models" is an example.

The statistical correlations that power this type of intelligent behaviour are extracted from datasets that are collected "from the wild" as recommended in [19]; finally, the current paradigm calls - or at least allows - for data annotation to be performed by using proxies or "implicit signals" such as the preferences revealed by user's activity, rather than explicit communicative acts directed at the agent.

Taken together these practices nearly deliver a "free lunch", since they do enable valuable predictions and decisions to be generated based on inexpensive data. Because of this, data was termed a "new natural resource". Based on these shortcuts, the field was able to deliver machine translation, speech recognition, object recognition, spell checkers, product recommendation, and much more. None of these tasks had been successfully automated under the previous AI paradigm.

These same shortcuts however are also behind some of the subtle side effects that started being observed: they encouraged a business model of assertive data collection, and personalised advertising; they allowed for subtle social biases to infiltrate elements of AI systems, such as word embeddings used to represent meaning; they might have encouraged the circulation of fake news by simply focusing on click through rate maximisation; they are behind the effect known as 'filter bubbles' [32]; they enabled delicate personal information to be inferred from publicly available information; they delegate potentially important decisions to systems that are not understandable to humans.

Taken together, these business and technical practices impose externalities on their users, by reducing their autonomy and possibly even violating some of their legal rights (such as the right to fair treatment, and more recently the right to an explanation). Redressing this will require significant and expensive re-work. It is this situation, of future costs being created, that we call the "ethical debt" of Artificial Intelligence.

Ethical debt, much like technical debt, results from taking shortcuts in the construction of software systems. Current practice in AI has made extensive use of this debt, and while this may well have been the correct decision to push the field forward, the time has come to start paying it back.

The Positive Side of Shortcuts. There can be situations where some debt is useful. Before we focus on some negative fall-outs of the current methodology, and on how to address them, we should briefly consider how the field would be if we had not made these choices. Finding a solution to AI's debt will require understanding what are the alternatives, and we should not fool ourselves into thinking that we have easy technical solutions: it is quite possible that solutions will have to come from different business models. While there seems to be some irony in the way we sought a free lunch and instead ended up in debt, and perhaps there is, the alternatives to the current path are not much better.

Without Shortcut 1 (machine learning instead of modeling), we would still

be struggling to come up with a set of explicit rules for a computer to distinguish a cat from a dog, or two handwritten letters, or to suggest a good book based on your previous reading history. That route had defied researchers for decades despite considerable investment, and there is no indication that it would work now. Without Shortcut 2 (gathering data already in the wild rather than producing our own) we would not only have had to prepay a significant amount of money to produce large amounts of training data, but we would have had to do this without even knowing if this would have been of any help. Furthermore, it is notoriously difficult to elicit certain types of information from human annotators, including rules to suggest books, or even just examples of appropriate suggestions for books or videos. Without Shortcut 3 (using implicit feedback to guide learning in agents) we would have been stuck with asking users to communicate to the agent explicitly and directly what type of items they wanted, and how satisfied they were with the suggestions they received. It has been known for a long time that users are often not fully aware of their information needs, and tend to resist providing this kind of feedback.

But at this point, with the knowledge and technology we have today, it has to be possible to reconsider some of those decisions.

Remedies to the Shortcuts. In the case of Shortcut 1, the use of causal (parametric, interpretable) models in certain domains might be mandated, even if accuracy might suffer, in the name of transparency of decisions. There are specific areas where users are entitled to explanations for consequential decisions, and it could be mandated that in these domains only weaker - but explainable - AI tools can be used. This would be a political decision and also a big change. Are we prepared to abandon black-box agents, paying the price of explicit modeling, and perhaps even hold back in certain areas where we fundamentally cannot develop those models? It seems unlikely, but we should have this conversation, at least for select sectors.

In the case of Shortcut 2 (training AI on data from the wild), we should at least be able to add some nuance: there can be types of data that can only be used for certain types of applications. Perhaps a given textual corpus can be suitable for training spelling correction agents, but not for learning the meaning of sensitive words (perhaps because it originates from a community with very different values than those that we want to be reflected in our agent). And a type of certification could even be imagined to state that origin. There are already specific lists of domains where decisions are expected to be unbiased, and for these domains we might request that AI agents are trained on better understood data sources, which may also be more expensive, making implicit biases explicit. We should care about our 'data supply chain' as much as we care about our food supply. This can be defined as the sequences of processes involved in the production and distribution of training data, which form the various models found in current AI systems. Each module might be based on different datasets, each of them in turn potentially shaped by yet other datasets. Are we prepared to pay the cost of generating, annotating and curating expensive datasets, matching the rigour used for clinical trial data? This might be unlikely, but we could regulate more closely the data markets, and develop screenings to

formalise which applications a given dataset is suitable for. Just gathering data from the wild without any further considerations is simply not safe.

In the case of Shortcut 3 (implicit feedback), it is possible to imagine that in certain domains the intelligent agent can only be allowed to learn from explicit, direct and voluntary communications from the user, rather than from observing their behaviour. This could be done in situations where there is the suspicion of filter bubbles or behavioural addiction. Deliberately using psychometric signals to infer how a user might react to a proposal might have to be banned, as possibly many forms of nudging. Regulating the use of implicit signals by intelligent agents seems to be a reasonable request

All this will probably cost more, might well reduce the performance of our systems, and their ease of use. Yet, domain by domain, we might decide that in some cases this is what we want. This would be part of paying back the ethical debt created over ten years ago by taking a series of shortcuts. We should not demonise those past decisions, as we would not have an AI industry today without them, but now the time has come to revisit some of them.

The implicit-explicit dimension. An important dimension that has so far been neglected in the analysis of social implications of AI systems, is the spectrum from implicit to explicit information. This covers the representation of knowledge within the agents; the mechanisms for making decisions; the signals contained in the training data; the biases that might be there; and the signals used as feedback to guide learning. This should also cover - of course - the explicit consent of users of AI systems; and explicit focus on "the data supply chain" as an object of analysis. Closer regulation of the data markets would be facilitated by using this distinction.

The use of explicitly created data might be mandated in situations where both the experimental conditions need to be tightly controlled, and the meaning of annotation must have been carefully agreed. The use of explicit relevance feedback by users would also be an important requirement in some conditions, preventing agents from 'eavesdropping' and using information unintentionally disclosed by users behaviour. Information about preferences is personal information, and gathering it requires explicit and informed consent, and explicit opt-in.

Explicit knowledge representation in AI models would allow easier sanity checks, both for bias and for explanations, but of course this is likely to limit the accuracy of systems. Explicit certification of biases that are present in the training data would also go a long way, allowing developers to make informed decisions about which components and data to include in their products. Future agents, in certain domains, might need to allow for the tracking of all sources of data which were used to train each of its components, and assessing their biases.

Finally, when we communicate with other users on social networks, or we access databases of content, it would be useful to drop the pretense that this is a direct interaction, making instead explicit the presence of an intelligent agent acting as intermediator- so that we can explicitly decide which engagement actions are communicative acts aimed at the other users, and which ones are

aimed at the recommending agent. We should not conflate the two types of communicative acts.

Acknowledgment. This work was partly supported by ERC Advanced Grant ThinkBIG.

References

[1] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[2] Justin Boyan, Dayne Freitag, and Thorsten Joachims. A machine learning architecture for optimizing web search engines. In *AAAI Workshop on Internet Based Information Systems*, pages 1–8, 1996.

[3] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederik Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

[4] James M Buchanan and William C Stubblebine. Externality. In *Classic papers in natural resource economics*, pages 138–154. Springer, 1962.

[5] Christopher Burr and Nello Cristianini. Can machines read our minds? *Minds and Machines*, pages 1–34, 2019.

[6] Christopher Burr, Nello Cristianini, and James Ladyman. An analysis of the interaction between intelligent software agents and human users. *Minds and machines*, 28(4):735–774, 2018.

[7] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[8] Nello Cristianini. Are we there yet? *Neural Networks*, 23(4):466–470, 2010.

[9] Nello Cristianini. On the current paradigm in artificial intelligence. *AI Communications*, 27(1):37–43, 2014.

[10] Nello Cristianini. A different way of thinking. *New Scientist*, 232(3101):39–43, 2016.

[11] Nello Cristianini. Intelligence reinvented. *New Scientist*, 232(3097):37–41, 2016.

[12] Nello Cristianini, James Ladyman, and Teresa Scantamburlo. Social machinery and intelligence.

[13] Nello Cristianini and Teresa Scantamburlo. On social machines for algorithmic regulation. *arXiv preprint arXiv:1904.13316*, 2019.

[14] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz S Kandola. On kernel-target alignment. In *Advances in neural information processing systems*, pages 367–373, 2002.

[15] W Cunningham. The wycash portfolio management system, experience report. *Proceedings on Object-oriented programming systems, languages, and applications (OOPSLA'92)*, 1992.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Ilias Flaounas, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital journalism*, 1(1):102–116, 2013.

[18] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

[19] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. 2009.

[20] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[21] Sen Jia and Nello Cristianini. Learning to classify gender from four million images. *Pattern Recognition Letters*, 58:35–41, 2015.

[22] Sen Jia, Thomas Lansdall-Welfare, Saatviga Sudhahar, Cynthia Carter, and Nello Cristianini. Women are seen more than heard in online newspapers. *PloS one*, 11(2):e0148434, 2016.

[23] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[24] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

[25] Thomas S Kuhn. The structure of scientific revolutions (3rd, 1996 ed.), 1962.

[26] Thomas Lansdall-Welfare, Stafford Lightman, and Nello Cristianini. Seasonal variation in antidepressant prescriptions, environmental light and web queries for seasonal affective disorder. *The British Journal of Psychiatry*, pages 1–4, 2019.

[27] Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465, 2017.

[28] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[29] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[31] Douglas W Oard, Jinmook Kim, et al. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems*, volume 83. WoUongong, 1998.

[32] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[33] Teresa Scantamburlo, Andrew Charlesworth, and Nello Cristianini. Machine decisions and human consequences. *arXiv preprint arXiv:1811.06747*, 2018.

[34] J Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166. ACM, 1999.

[35] Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. Biased embeddings from wild data: Measuring, understanding and removing. In *International Symposium on Intelligent Data Analysis*, pages 328–339. Springer, 2018.