Evolving AI from Research to Real Life - Some Challenges and Suggestions

Sandya Mannarswamy[1] and **Shourya Roy**[2]
[1] Conduent Labs India
[2] American Express Big Data Labs
sandyasm@gmail.com, shourya.roy@aexp.com

Abstract

Artificial Intelligence (AI) has come a long way from the stages of being just scientific fiction or academic research curiosity to a point, where it is poised to impact human life significantly. AI driven applications such as autonomous vehicles, medical diagnostics, conversational agents etc. are becoming a reality. In this position paper, we argue that there are certain challenges AI still needs to overcome in its evolution from *research to real life*. We outline some of these challenges and our suggestions to address them. We provide pointers to similar issues and their resolutions in disciplines such as psychology and medicine from which AI community can leverage the learning. More importantly, this paper is intended to focus the attention of AI research community on translating AI research efforts into real world deployments.

1 Introduction

It was the best of times,
it was the worst of times,
it was the age of wisdom,
it was the age of foolishness,
it was the epoch of belief,
it was the epoch of incredulity,
it was the spring of hope,
it was the winter of despair.

(Charles Dickens)

There is almost a Dickensian feel to the current state of the evolution of Artificial Intelligence (AI) in spite of tremendous progress in the field. Crossing the various stages of science fiction imagination and academic research curiosity, AI at its current trajectory of evolution, has already started impacting many of the aspects of the modern human life [Stone *et al.*, 2016]. Driverless cars are on the road without any fallback human driver on-board[1]; AI-based medical diagnosis software is performing better than certified doctors which is now recognized and published in leading medical journals [Gulshan *et al.*, 2016]; conversational agents are becoming more and more pervasive in all spheres of lives [Venkatesh *et al.*, 2018]; AI-enabled agents are beating world champions in various complex games [Silver *et al.*, 2017]. Riding on advances in basic research and such practical demonstrations in real-life settings, Artificial General Intelligence (AGI) seems within reach of mankind soon.

However, the re-emergence of AI from the long AI winter [Hendler, 2008] has been marked by both hype and hope [Brooks, 2017]. Scientific progress in AI has been explosive, with research publications being churned out at a rapid pace [Mauro, 2017]. This has been accompanied by considerable industrial efforts bringing the research into real world deployment and monetization. Among them neural networks and Deep Learning (DL) techniques have been leading the race with almost every AI conferences are swamped by advances in deep learning related papers. The current *AI Spring* has been characterized by a greater emphasis on empiricism, with experimentation driving research publications to a large extent. Advancements in computing paradigms have enabled faster iteration and experimentations, while a matching progress in theoretical underpinnings of the field taking a back seat. As publications mushroom in the race towards state of the art numbers in various AI tasks, there has been an increasing apprehension that the scientific rigor is perhaps getting sacrificed at the altar of too rapid innovation [Rahimi and Recht, 2017].

The growth of AI research and its real-life applications has been greeted with speculations and concerns from different spheres. Back and forth arguments between the believers and the naysayers of AI and DL have been on forefront of social media almost daily. In this paper, we argue that while AI has made significant technological advances, there are certain inherent issues which require systematic redressal, as it is at the cusp of transformation from research to real life. These issues include:

- **Explainability**: As AI is increasingly penetrating into real-life applications, decisions made by AI-enabled systems need to be explainable. Why a loan applicant was rejected? On what basis a patient was diagnosed positive for a malignant disease? The ability to provide justifiable and reliable evidences for their decisions would certainly increase the trust of users.

[1] https://tinyurl.com/y83wmu6z

- **Fragility**: AI/DL systems in various settings have occasionally been reported to be functioning in surprising and undesirable manners when put in a different but related environment. With very minor changes in the input data, as minor as one-pixel in an image, can lead to very different output [Su *et al.*, 2017] which may have disastrous implications for real-life applications based on deep computer vision models. On the research side, adversarial attacks and heavy reliance on fine tuning of large number of parameters question the applicability of many research work in real-life applications.
- **Research Practices**: While sharing of code and implementation details have increased recent times, reproducibility of experiments from research papers have been a challenge. Tall claims basis of empirical evaluation from papers have often been found to be difficult to reproduce; partially owing to increasing complexity of models, number of hyperparameters etc. With the growing flag-planting practice in pre-print services such as arxiv[2], researchers are racing against time to often submit their half-baked work. In addition, publication bias towards positive results have not been serving the community as a whole.

A scientific discipline encounters different challenges as it evolves and matures. In its earlier evolution, challenges around its inability to demonstrate its promised impact on real life applications [Lighthill, 1973] was responsible for the long *AI winter* a few decades back. In its current incarnation, AI has clearly demonstrated its impact with many real world applications but going forward it needs to overcome the barriers/challenges for AI research to make the transition from being a 'cool technology' to a 'real life presence' in human society.

Some of the above-mentioned challenges such as *explainability* and *fragility* are unique to AI field itself, while other challenges such as *research practices* are common across other scientific disciplines as well. In particular, some of these challenges have been earlier encountered in the evolution of the related disciplines of psychology and medicine. Research community in those disciplines have come up with novel methods to address these issues. For instance, the field of psychology came up with "Registered Reports" mechanism [Editorial, 2017]. A Registered Report is a form of scientific publication in which methods and proposed analyses are pre-registered and peer-reviewed prior to research being conducted (referred to as stage 1). Research protocols specified in stage 1 are provisionally accepted for publication before data collection and experimentation commences. Once the study is completed, the author will submit the completed article including results and discussions sections (referred to as stage 2). Provided the study adhered to the agreed upon protocols, the article is accepted for publication. This can help to avoid reviewers' bias towards the experimental improvements without giving adequate weightage for novelty and correctness. Controlled research deployment in terms of phased clinical trials [Sedgwick, 2014] has helped mitigate the risk of "unknown unknowns" in pharma field. Similarly the role of federated approving authority such as Food and Drug Administration (FDA) in the United States may be relevant for AI deployments which can impact human safety. We argue that AI community needs to leverage many of these inter-disciplinary learnings as it evolves from being a cloistered academic research to a major presence in human activities. In this paper, we outline how AI can leverage some of the learnings from related disciplines to address the above cited challenges.

Scientific research in any discipline needs to be self-correcting for it to sustain its rapid pace of real progress instead of falling victim to short-term shallow victories. We opine that AI in its evolution is currently at an inflection point where the community has to retrospect about these challenges, self-criticize and if needed, self-correct itself, which requires community effort. This position paper is an initial step in fostering community efforts towards understanding and addressing the challenges in taking AI from research bench to real life.

This paper is obviously not the first attempt to identify the some of the challenges to a successful *translational AI*. While there has been a growing concern among the AI community about these issues, the discussions have largely been in silos. We believe that these issues in translational AI are inter-related and hence there is a need for a concerted community effort to address them. Addressing these challenges may also require a greater degree of control and regulations. Hence there is a pressing need for the AI research community to self-devise appropriate measures of consensual control without impacting scientific progress. This paper argues that leveraging the learnings from the related disciplines of psychology and medicine may help in that direction.

This paper is organized as follows. In the next section, we list the challenges that AI is facing as it evolves from research to real life. In Section 3, we discuss how related disciplines such as psychology and medicine have successfully coped with some of these challenges and how AI community can leverage these learnings. We conclude the paper in section 4 briefly highlighting some of the future directions.

2 Challenges in Evolving AI from *Research to Real Life*

As AI makes the transition from "research bench to real life", the challenges faced by it, can be broadly categorized into issues inherent to AI and issues non-inherent to AI. We did not intend to be exhaustive in our attempt to list the challenges. Rather we focus on specific challenges, which we believe, would benefit from discussion and attention by the larger AI community. We first discuss the AI specific challenges.

2.1 AI Specific Challenges

AI applications which directly impact human life and safety such as driverless cars and AI driven clinical decision-making systems, are now starting to emerge. Deploying AI at scale in such complex scenarios also brings with it, certain key requirements. For instance, consider the task of clinical decision-making. It is expected that,

[2]https://arxiv.org/

- the human doctors can explain the rationale behind their decisions when needed (*interpretability*),
- they will make the right decisions for the patient, even if the certain symptoms of the patient are deliberately trying to mislead them (*adversarial vulnerability*),
- their clinical decisions will not change due to changes in hospital setup (*anti-fragility*)
- they can still make the right decisions if they shift their practice from Ohio to California (*transferability*).
- other doctors with similar expertise can replicate their decisions when presented with the same evidence (*reproducibility*).

We expect similar high standards in AI applications operating under such scenarios. Unfortunately, the current state of AI driven applications are far from this desired state.

Interpretability

In this seminal article on interpretability, Lipton opined that interpretability of machine learning models is an ill-defined concept having various desiderata including trust (*is a model trustworthy*), robustness (*is a model behaving consistently*) to transparency (*is a model's working understandable*) [Lipton, 2016]. He argued that it is important to narrow down on the interpretation of interpretability of AI systems, before trying to address the challenges. Early days of statistical machine learning systems (in the 90s and 2000s), hand-designed features used to be prevalent for developing models. While the practice of designing features was a time consuming and expertise-seeking task, the features used to be helpful for understanding predictions from learning algorithms. However with the growing popularity of deep learning, interpretability of techniques have taken a backseat. Deep learning models while achieving state of the art numbers across various modalities viz. text, image, video, audio etc., have become opaque and less transparent about their functioning. Kenneth Church [Church, 2017] neatly summed this up when he said:

> *We are seeing lots of papers, these days, that say, I did it, I did it, I did it, but I don't know how! Now even the author of the paper doesn't know how the machine does what it does.*
>
> (Kenneth Church)

As of now, the way at which an AI system arrives at a particular decision continues to remain as a black box to us. Owing to lack of interpretability of predictions, machine learning models have been accused of being unfair and biased. In the book "Weapon of Math Destruction", author O'Neill brought forward a number of examples of how machine learning models are biased in making predictions for real-life use-cases from financial loan approval to crime prediction in big cities etc [O'Neil, 2016].

Researchers have realized the importance of interpretability and there have been a number of work published in recent times towards better explainability of neural deep models [Olah *et al.*, 2018]. Visualization of deep networks in terms of activations of hidden layers to explain how it detects what it detects, has been gaining tremendous popularity in computer vision. Ablation studies and heatmaps are common in natural language processing techniques towards identifying attribution of different components of deep neural architecture [Ding *et al.*, 2017]. With increasing complexity of models and architectures, the ability of explaining predictions considering model as a black-box [Koh and Liang, 2017] is gaining attention.

While there is much work in progress towards improving interpretability [Olah *et al.*, 2018], the ideal state of having explainable, evidence driven AI decisions still remains a challenge. The inherent challenge is the fact that there are two separate spaces wherein humans and current AI systems operate. Humans reason in terms of concepts/symbolic space of world, whereas most of the current AI/ML systems operate in a mathematical space of numbers onto which human specified information has been transformed into. Any explanatory interface for AI decisions needs to map across these two symbolic spaces, which can potentially lead to lossy transmission of information. It may be worthwhile to recall Polanyi's paradox [Polanyi, 1966] in this context - "We often know more than we can tell". We hypothesize that Polanyi's paradox may prove to be equally true for AI systems and hence the goal of 100% interpretability of AI systems may not be practically realizable. We note that the goal of full interpretability can make AI systems more vulnerable to adversarial attacks and hence carefully designed access mechanisms for interpretable explanations may need to be devised. This lends weight to our argument that AI evolution challenges needs to be considered holistically and not in silos.

Fragility

Given the astounding successes AI applications have had in recent years, it is easy to fall into the trap of misinterpreting what AI models do and overestimating their abilities. For example, if a model can perform the task of image captioning accurately, it does not imply that the model understands the meaning/concepts that gave rise to these images. This can lead to unpleasant surprises when slight perturbation in the data can lead to very inaccurate results. This has been demonstrated time and time through adversarial vulnerability exploits against many of the deep learning systems in text/vision/speech. For instance, Su et al. [Su *et al.*, 2017] showed that 70.97% of the natural images can be misclassified by DNNs by modifying just one pixel with 97.47% confidence on average. Further Ilyas et al. [Ilyas *et al.*, 2017] showed that black-box driven adversarial examples can be successfully employed against a commercially deployed machine learning system. AI applications in other modalities also display such vulnerability. Recently Jo and Bengio [Jo and Bengio, 2017] demonstrated that deep neural networks have a tendency to learn surface statistical regularities as opposed to high level abstractions. This underscores the point that current crop of AI/DL approaches are not learning higher level semantic relations between concepts, instead their performance is possibly due to their picking up the superficial statistical cues present in both the train and test data.

Besides many examples in the context of computer vision, state of art models in NLP are also vulnerable to minor perturbations of the data. Jia et al. [Jia and Liang, 2017] showed that

many of the models in reading comprehension test fail with simple adversarial examples, because they can't distinguish a sentence which answers the question, from one which merely contains words common with it. Many recent studies have shown that simple examples produced either by human or machine, can cause AI/DL systems to fail spectacularly [Yuan *et al.*, 2017]. For example, a recently released textual entailment demo was criticized on social media for predicting textual entailment with 92% confidence: "John killed Mary" → "Mary killed John"[3]. In addition to adversarial vulnerability, AI driven systems are also extremely fragile to experimental setup such as hyperparameters, initialization and many of the claimed performance improvements can be attributed purely to such setup variations [Lucic *et al.*, 2017]. Current AI/ML systems are also highly specific for the task they are designed for and the data they earn from. Their learnings do not transfer well across tasks and across domains.

Such lack of transferability is evident not only in the field of research but also in practical and real-life settings. Last year, an implementation of IBM Watson [Chen *et al.*, 2016] at University of Texas MD Anderson Cancer Center reportedly failed resulting into a loss of $62 million [Mulcahy, 2017]. The project was to build an Oncology Expert Advisor (OEA) by integrating with Electronic Medical Records (EMR). While many reasons have been cited for Watson to live up to the Jeopardy fame in real-life, the most prominent one being Watson's dependence on task and associated corpus for every new task. It required months of training and well-formatted data ingestion to be able to draw any conclusion in any task This dramatically limits its use in clinical settings.

Limited Generalization - The Root of All Evils?

While research is in progress to address these challenges, it is important to note that, the current AI/DL systems have inherently limited generalization capability [Marcus, 2018; Chollet, 2017]. They map human world symbols (concepts) into mathematical symbols (vectors), perform geometric transformations on them to learn a relation to the output label without any understanding the meaning of the input abstractions. Unlike humans who can build complex abstract models of their current situations and postulate different non-experienced scenarios from it, Current AI systems are only capable of dealing with situations/data that are very close/similar to earlier seen data/experiences. Their poor transferability, fragility and adversarial vulnerability are a direct result of this lack of understanding of higher level abstractions/concepts (unlike human thinking, which can perform extensive generalization instead of local generalization). While there is work in progress towards AI systems which can overcome these challenges [Zhang *et al.*, 2016; Neyshabur *et al.*, 2017; Kawaguchi *et al.*, 2017; Nichol *et al.*, 2018], it is responsibility of us, as AI practioners to ensure that deployment of AI in real life situations are carefully analyzed for such vulnerabilities, and augmented with suitable safeguards and human oversight wherever appropriate.

[3]https://newgeneralization.github.io/

2.2 Other Challenges

In addition to the above challenges inherent to its domain, AI also faces certain common issues faced by other scientific disciplines such as psychology and medicine, in translating research to real life. In this section, we list a few of these challenges and outline how AI can leverage the learnings from the other disciplines.

Challenges Associated with Research Publications

These include poor reproducibility, lack of support for publication of negative results and publication bias towards positive results [Sculley *et al.*, 2018]. We first discuss the issue of reproducibility. While the terms *replicate* and *reproduce* are often used interchangeably, we use the term replication for repeating the same experiments mentioned in the paper on the same data, while reproducing is to repeat the technical approach in principle on same/different data and see if conclusions hold true.

One solution widely proposed as a panacea to the reproducibility challenge in AI research community is the push towards sharing the code and data used in the research experiments, so that the experimental results can be validated independently. It was recently reported in a survey on reproducibility in AI that of 400 recent research papers published in top AI conferences, only 6% had provided the code behind the techniques [Hutson, 2018; Gundersen and Kjensmo, 2018].

While sharing code is an important step, this does not address the problem in its entirety. Many a time, the shared code does not produce reported results or even does not run. Also there is no incentive for reproducing experimental results. With the prevalence of the "Publish or Perish" research culture, who will own the burden of replicating or reproducing experiments when they do not yield to any publications? Some of these issues have been pointed out in [Sculley *et al.*, 2018]. While there has been some effort recently to actively encourage reproducing the results of published research papers through the NIPS 2017 paper implementation challenge[4], such efforts have been few and far between.

While one may be tempted to think that replication efforts typically occur when researchers use a prior work as baseline, such replication efforts are typically colored by the researchers trying to showcase their own work in best light (use the best hyper-parameter setting) against the baseline (just run with any set of hyper-parameters). Another constraining factor is the need for considerable computing resources that may be needed to replicate an experiment, which can act as deterrent. Given these constraints, we argue that current solution of requiring researchers to publish code (and where possible, data) does not completely address this challenge. While efforts such as OpenML [Vanschoren *et al.*, 2014] and ParlAI [Miller *et al.*, 2017] have made positive steps in direction, AI community needs to devise novel mechanisms to address this issue. We believe that existing efforts at replication primarily fail because of

- Lack of intellectual/monetary incentive for reproducibility/replication experiments.

[4] https://nurture.ai/nips-challenge

- No formal mechanism to verify shared code is runnable and replicates results claimed in the paper. While sharing code counts positively for getting published, reviewers don't have the bandwidth to verify that the shared code performs what it sets out to do, except in specific competition/task settings.

We argue that AI community is better-placed than other disciplines in addressing this issue because we can use AI itself to attack this problem. We need to design AI solutions/tools which can perform this task of verification. Given a research paper which contains the textual details of the technique, the pointer to the source code and the data, as well as the experimental numbers claimed, we argue that tools can be designed which can mitigate the replication workload. A recent paper by [Sethi *et al.*, 2018] proposed generating neural network code automatically based on the diagrams provided in the research paper. More such tools are needed to address the replication crisis in AI.

Other related challenges include lack of support for negative results and publication bias towards positive results. We can understand the magnitude of the problem felt by AI research community as expressed in the following quotes from one of the leading NLP researchers:

> *"Recently at ACL conferences, there has been an overfocus on numbers, on beating the state of the art. Call it playing the Kaggle game. More of the field's effort should go into problems, approaches, and architectures."*
>
> (Chris Manning [Manning, 2015])

Challenges in Deployment

Translating basic AI research findings into applications deployed in real life is a long-drawn and cumbersome task. While AI has crossed the chasm in demonstrating its impact in niche real life applications, much of its potential for real life impact is yet to be realized. Hence it requires that this translational cycle needs to be carefully nurtured and directed by the AI research community to avoid the recurrence of another AI winter. There are a multitude of deployment challenges such as safety, legality, cultural impact, auditability, explainability, fairness in deployment scenarios. There is work in progress towards defining accountability for AI decision making algorithms that impact real life [Diakopoulos, 2016]. Algorithmic Impact Assessments have been proposed as a framework in this direction which is a promising step in addressing some of the critical AI deployment issues [A.Campolo, 2017].

3 Leveraging Learnings from Related Disciplines

In this section, we outline how learnings from disciplines such as psychology and medicine be leveraged in AI in addressing the above mentioned challenges.

3.1 Addressing Research Publications Issue

Issues such as poor reproducibility, publication bias towards positive results are not unique to AI. Various other scientific disciplines such as psychology, biology and medicine have also faced similar issues [Munafò *et al.*, 2017]. In fact, the reproducibility crisis in medicine is still far from over. Medicine's reproducibility crisis is well documented in the work of John Ioannidis [Ioannidis, 2005]. In a detailed analysis of 45 most highly cited papers in medicine, he found that 16% of them were contradicted and another 16% of them were found to have over-estimated their impacts. His 2005 paper "Why Most Published Research Findings Are False" showed that such failures are to be expected, given that majority of published findings are not based on large, randomized-controlled trials, but small observational studies or studies with small population sizes. This is true for any emerging scientific discipline on its evolution from research to real life, such as AI as well. Since initial research findings are with single/small data sets or specific experimental setups, we need to expect that many of these research findings may not be reproducible with larger studies. However this requires that reproducibility needs to be encouraged actively and replication studies should become the norm instead of being rarity. Instead of judging research publications merely based on performance numbers, greater weight should be given to novelty of the technique, underlying theory and rigorousness of the experimentation protocol.

One formal mechanism to deal with this issue is the concept of "Registered Reports", originally proposed in Psychology and later adopted in other related disciplines as well [Editorial, 2017]. A Registered Report is a mechanism wherein the research publication process is split into two stages. In the first stage, the experimental hypothesis, the technique for validating the hypothesis, the experimental protocol etc. are recorded in a paper and registered for publication. This document is peer-reviewed and the decision to publish is made on the basis of this pre-registered report (without access to any experimental results). The researcher then carries out the experimentation as per the protocol and the results, whether positive or negative, gets published. Since the experimentation protocol is clearly stated, it aids in replication effort. Registered Reports help to address publication bias towards positive results, since the publication decision is determined without access to experimental results. This also helps to mitigate the problem of not reporting negative results.

We believe that the time has come for AI research community to adopt Registered Reports. It is true that Registered Reports can delay publication time, and not all research may not be suitable for Registered Reports format. Hence a phased adoption of Registered Reports mechanism can be appropriate for AI research. We also recommend that AI research community should support replication studies and negative results publication as separate tracks in major conferences and journals. The AI community should also introduce a stringent retraction watch mechanism in cases where multiple independent replication attempts of major findings fail.

3.2 Addressing Deployment Challenges

We believe that AI can leverage learnings from translational medicine ("the bench to bedside initiative"). The key principle of translational medicine is the "Precautionary Principle (PP)" which states that if an action has a potential risk of causing harm to the public domain, the action should not be taken

in the absence of scientific near-certainty about its safety. Just as in translational medicine, given the risks and uncertainity in complex situations where AI can augment/replace human decision making, Precautionary Principle definitely needs to be kept in mind for mitigating social impact risk in translational AI. Medicine has gone through a complex and difficult trajectory in its evolution from research bench to real life. Translational Medicine came up with the driving principle of "carefully controlled deployment" of its research findings into clinical practice. Controlled deployment has been realized through staged translation of research [Woolf, 2008]. Translational medicine is typically staged into 4 stages T1-T4. A good example of staged translation is vaccine development. T1 phase consists of basic vaccine research in lab, with animals and limited human trials. At the T2 (lab to clinical trial) stage, highly controlled trials are run to determine safety and efficacy of the new vaccine. At T3 level (from clinical trial to practice), US FDA authorizes the general use of vaccine. FDA and US Centers for Disease Control and Prevention (CDC) would monitor adverse reactions through their existing reporting system. At T4 level (practice to policy), CDC would evaluate data and create a policy for use in clinical and community settings. Such staged translation of research finding (a new vaccine) enables improved patient outcomes while mitigating risk.

We argue that translational AI would benefit from emulating such controlled deployment of technology especially in cases of complex scenarios wherein human decision making is being augmented/replaced by AI (an example could be driverless cars). While the translational cycle can be quite different for AI (certain phases can be omitted/refined), it is essential that any applications of AI replacing human decision making needs to have carefully controlled release, as in case of translational medicine. It is essential that AI community devises appropriate guidelines for achieving this controlled deployment of its technology.

Translational medicine has also put in place, effective measures for improving societal benefit while mitigating risk. These measures include:

- Evidence Based Medicine [Guyatt *et al.*, 1992] – Clinical decision making should only be guided by research findings with strongest grade of evidence (meta-analyses, systematic reviews, and randomized controlled trials).
- Comprehensive Health Technology Assessment (HTA) [Luce *et al.*, 2010]
- Outcomes based research including patient reported outcomes [Hickok, 1997]
- Clinical effectiveness and comparative effectiveness research [Greenfield and Rich, 2012]
- Multi-stage regulatory mechanism for approvals through an industry independent regulatory agency such as FDA
- Effective post-deployment surveillance reporting and adverse events reporting [Rizwanuddin, 2003]

We argue that many of these mechanisms with suitable modifications need to considered for AI deployment in society to mitigate the unknown risks. In particular we suggest that

- AI community should constitute a 'translational AI' workforce, whose focus would be on translating AI research into real life.
- Translational AI needs to include experts from translational medicine and translational psychology to cross-leverage learnings from those disciplines
- Real life large scale AI deployments should follow a process of controlled release (phased trials). Depending on the field of applications (medical AI/legal AI/financial AI), different mechanisms would need to be evolved.
- AI research community needs to put in place, a federated deployment advisory board consisting of both academic and industry, which should review large scale AI deployments. While one may argue that each of the application fields (medical AI/legal AI) have their own regulatory mechanism, these domain specific regulatory authorities may not have adequate AI deployment expertise and the use of AI community driven regulatory authority is to aid these domain specific authorities and not to supplant them.
- Comprehensive AI Technology Assessments (similar to Health Technology Assessments) need to be carried out before deploying a particular AI application in field. (While we note that not all AI applications need such careful control, we need mechanisms to classify AI applications appropriately (similar to Over the Counter Drugs vs Scheduled Drugs) into classes which need varying degrees of control and regulation based on their societal impact and human risk.
- Just as translational medicine benefitted from Patient Reported Outcomes and Informed Patient Movement [R Deshpande *et al.*, 2011; Parker, 2006], translational AI also needs to actively include its end consumers of the technology in its processes.
- AI needs its own Cochrane Report [Chalmers *et al.*, 1992] for evidence based translation of research to practice!
- We note that any measures of control and regulation needs to be devised, without stifling innovation and enabling faster translation. Here again, some of the learnings from translational medicine would be effective for us to leverage. A case in point is FDA's Critical Path Initiative [5] in accelerating translational medicine.

4 Conclusion

In this paper, we brought forward a few challenges that AI still needs to overcome in its evolution from *research to real life* as it demonstrates significant potential in impacting human lives. For a subset of the challenges we suggested certain mechanisms to deal with those, piggybacking on learnings from other fields such as psychology and medicine. We did not intend to be exhaustive and are cognisant of several other challenges which require due attention such as safety, ethical

[5] https://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/default.htm

and cultural issues etc. which can be taken up in future. Finally, we hope that the paper would stimulate AI researchers and practitioners' thoughts towards addressing the challenges listed as well as facilitate productive discussions in the community.

References

[A.Campolo, 2017] A.Campolo. AI Now 2017 Report. pages 1–4, 2017. https://ainowinstitute.org/AI_Now_2017_Report.pdf.

[Brooks, 2017] Rodney Brooks. The seven deadly sins of predicting the future of AI. *MIT Technology Review*, pages 1–4, 2017. https://tinyurl.com/y7z5wrl9.

[Chalmers *et al.*, 1992] Iain Chalmers, Kay Dickersin, and Thierry Coquand Chalmers. Getting to grips with Archie Cochrane's agenda. *BMJ*, 305 6857:786–8, 1992.

[Chen *et al.*, 2016] Ying Chen, JD Elenee Argentinis, and Griff Weber. IBM Watson: How cognitive computing can be applied to life sciences research. *Clinical Therapeutics*, 38:688–701, 2016.

[Chollet, 2017] Francois Chollet. *Deep Learning with Python*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2017.

[Church, 2017] Kenneth Ward Church. Emerging trends: I did it, I did it, I did it, but. . . *Natural Language Engineering*, 23(3):473–480, 2017.

[Diakopoulos, 2016] Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, January 2016.

[Ding *et al.*, 2017] Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1150–1159, 2017.

[Editorial, 2017] Editorial. Promoting reproducibility with registered reports. *Nature Human Behavior*, pages 1–2, 2017. https://www.nature.com/articles/s41562-016-0034.

[Greenfield and Rich, 2012] Sheldon Greenfield and Eugene C. Rich. Welcome to the journal of comparative effectiveness research. *Journal of comparative effectiveness research*, 1:1–3, 2012.

[Gulshan *et al.*, 2016] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316 22:2402–2410, 2016.

[Gundersen and Kjensmo, 2018] Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1–8, 2018.

[Guyatt *et al.*, 1992] Gordon Guyatt, John Cairns, and David Churchill. Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA*, 268(17):2420–2425, 1992.

[Hendler, 2008] James Hendler. Avoiding another AI winter. *Intelligent Systems, IEEE*, 23:2–4, 04 2008.

[Hickok, 1997] Durlin E. Hickok. Outcomes-based research: What is it and how do we do it? *Seminars in Perinatology*, 21(6):467 – 471, 1997.

[Hutson, 2018] Matthew Hutson. Missing data hinder replication of artificial intelligence studies. *Science*, pages 1–4, 2018.

[Ilyas *et al.*, 2017] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Query-efficient black-box adversarial examples. *CoRR*, abs/1712.07113:1–8, 2017.

[Ioannidis, 2005] John P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[Jia and Liang, 2017] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics, 2017.

[Jo and Bengio, 2017] Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. *CoRR*, abs/1711.11561:1–8, 2017.

[Kawaguchi *et al.*, 2017] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *CoRR*, abs/1710.05468:1–8, 2017.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1885–1894, 2017.

[Lighthill, 1973] James Lighthill. Artificial intelligence: A general survey. In *Artificial Intelligence: a paper symposium*, pages 1–21, 1973.

[Lipton, 2016] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490:1–8, 2016.

[Luce *et al.*, 2010] Bryan Luce, Michael Drummond, Bengt JÃ¶nsson, Peter Neumann, J Sanford Schwartz, Uwe Siebert, and Sean D Sullivan. EBM, HTA, and CER: clearing the confusion. *The Milbank quarterly*, 88:256–76, 06 2010.

[Lucic *et al.*, 2017] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs Created Equal? A Large-Scale Study. *ArXiv e-prints*, pages 1–8, 2017.

[Manning, 2015] Christopher D. Manning. Computational linguistics and deep learning. *Computational Linguistics*, 41:701–707, 2015.

[Marcus, 2018] Gary Marcus. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631:1–27, 2018.

[Mauro, 2017] Gordon Mauro. Six graphs to understand the state of AI academic research. pages 1–4, 2017. https://blog.ai-academy.com/.

[Miller *et al.*, 2017] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84. Association for Computational Linguistics, 2017.

[Mulcahy, 2017] Nick Mulcahy. Big Data bust: MD Anderson-Watson project dies. *Medscape*, pages 1–3, 2017.

[Munafò *et al.*, 2017] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021+, 2017.

[Neyshabur *et al.*, 2017] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30*, pages 5947–5956, 2017.

[Nichol *et al.*, 2018] Alex Nichol, Vicki Pfau, Christopher Hesse, Oleg Klimov, and John Schulman. Gotta learn fast: A new benchmark for generalization in RL. *CoRR*, abs/1804.03720, 2018.

[Olah *et al.*, 2018] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, pages 1–8, 2018. https://distill.pub/2018/building-blocks.

[O'Neil, 2016] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY, USA, 2016.

[Parker, 2006] Ruth M. Parker. What an informed patient means for the future of healthcare. *PharmacoEconomics*, pages 29–33, 2006.

[Polanyi, 1966] Michael Polanyi. *The Tacit Dimension*. Routledge and Kegan Paul, 1966.

[R Deshpande *et al.*, 2011] Prasanna R Deshpande, Surulivel Rajan, B Lakshmi Sudeepthi, and C P Abdul Nazir. Patient-Reported Outcomes: A new era in clinical research. 2:137–44, 10 2011.

[Rahimi and Recht, 2017] Ali Rahimi and Ben Recht. Reflections on random kitchen sinks. pages 1–3, 2017. http://www.argmin.net/2017/12/05/kitchen-sinks/.

[Rizwanuddin, 2003] Ahmad Syed Rizwanuddin. Adverse drug event monitoring at the Food and Drug Administration. *Journal of General Internal Medicine*, 18(1):57–60, 2003.

[Sculley *et al.*, 2018] D Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. pages 1–4, 2018. https://openreview.net/forum?id=rJWF0Fywf.

[Sedgwick, 2014] Philip Sedgwick. What are the four phases of clinical research trials? 348:1–4, 2014.

[Sethi *et al.*, 2018] Akshay Sethi, Anush Sankaran, Naveen Panwar, Shreya Khare, and Senthil Mani. DLPaper2code: Auto-generation of code from deep learning research papers. *CoRR*, abs/1711.03543:1–8, 2018.

[Silver *et al.*, 2017] David Silver, Julian Schrittwieser, and et al. Mastering the game of Go without human knowledge. *Nature*, 2017.

[Stone *et al.*, 2016] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. Ai and life in 2030. Technical report, 2016. http://ai100.stanford.edu/2016-report.

[Su *et al.*, 2017] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *CoRR*, abs/1710.08864:1–8, 2017.

[Vanschoren *et al.*, 2014] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Exploration Newsletter*, 15(2):49–60, June 2014.

[Venkatesh *et al.*, 2018] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. On evaluating and comparing conversational agents. *CoRR*, abs/1801.03625:1–10, 2018.

[Woolf, 2008] Stephen Woolf. The meaning of translational research and why it matters. *JAMA*, 299(2):211–213, 2008.

[Yuan *et al.*, 2017] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *CoRR*, abs/1712.07107:1–8, 2017.

[Zhang *et al.*, 2016] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.