NICE: Four Human-Centered AI Principles for Bridging the AI-to-Clinic Translational Gap

Julia Gong*
jxgong@cs.stanford.edu
Stanford University
Stanford, California, USA

Rebecca Currano
Stanford University
Stanford, California, USA

David Sirkin
Stanford University
Stanford, California, USA

Serena Yeung
Stanford University
Stanford, California, USA

F. Christopher Holsinger
Stanford University
Stanford, California, USA

ABSTRACT

Despite the rapid development and application of artificial intelligence (AI) methods to myriad challenges of clinical interest in recent years, a strikingly low proportion is eventually truly deployed in the healthcare system. In this paper, I propose a design-inspired framework of four pillars—Needs, Iteration, Collaboration, Error Prevention (NICE)—for closing this translational gap by examining the literature, my own research, and clinical collaborations. Focusing on building clinician-centered methods for in-clinic care, I outline (1) several key factors that fuel the translational gap between AI methods and clinical implementation, (2) how these factors can ideally be addressed in the long run, and most critically, (3) exciting directions that AI researchers and engineers can begin taking immediately to play our part in advancing these goals.

KEYWORDS

AI for healthcare, human-centered AI, patient-centered care, translational research, design principles

ACM Reference Format:
Julia Gong, Rebecca Currano, David Sirkin, Serena Yeung, and F. Christopher Holsinger. 2021. NICE: Four Human-Centered AI Principles for Bridging the AI-to-Clinic Translational Gap. In *Virtual '21: ACM CHI Workshop on Realizing AI in Healthcare: Challenges Appearing in the Wild, May 08–09, 2021, Virtual.* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Recent years have seen the rapid development and application of artificial intelligence (AI) methods to myriad challenges of clinical interest. However, despite the great potential of these works, a strikingly low proportion is eventually truly deployed in the healthcare system. As a student researcher whose work lies precisely at the crossroads of machine learning and healthcare and has deep collaborations with clinical partners who express similar concerns, I have been invested in understanding the mechanisms behind this disparity. Given this translational gap between AI methods and the clinic, what can we do to bridge it?

Drawing on the literature and my own experiences, this paper hopes to seize this moment of collective reflection in the AI and medical communities as an opportunity to propose a design-inspired framework for explaining and addressing this gap, including actionable guidelines for researchers and engineers in my home field of AI. Inspired by my medical collaborators in head and neck surgery, I will focus my discussion on building *clinician-centered methods for in-clinic care*, where clinicians are the end users.

Below, I outline (1) several key factors that appear to fuel this translational gap between AI methods and clinical implementation, (2) how these factors can ideally be addressed in the long run, and most critically, (3) exciting directions that AI researchers and engineers can begin taking immediately to play our part in moving the needle toward these goals.

2 THE NICE FRAMEWORK

This framework proposes four principles as key pillars of translational medical AI research, each drawing inspiration from the design process: **N**eeds, **I**teration, **C**ollaboration[1], and **E**rror Prevention and Monitoring (NICE). Weaknesses in each pillar correspond to root causes that contribute to the AI-to-clinic translational gap, which in turn inform long-term goals and actionable proposals for AI researchers and engineers.

2.1 Needs-Focused Innovation

At the root of designing useful tools is the concept of fulfilling a true *need*. Dr. Robert McKim, a founding figure in Stanford's human-centered design program, advocated that all designs arise from human needs, which arise organically from their contexts [39]. This principle that [3] referred to as the Need-Design Response likewise applies to the medical AI context. Some forward-thinking health technology organizations like Stanford's Center for Biodesign specifically emphasize the importance of "collect[ing] hundreds of needs" in the early stages of the biodesign process [8].

Indeed, general AI research communities have long upheld a similar principle that new work in a field should identify problems in prior work and compellingly illustrate how a proposed solution

*This is a position paper by J. Gong, which incorporated feedback from her AI and clinical research advisors, S. Yeung and F. C. Holsinger, as well as R. Currano and D. Sirkin from the Stanford Center for Design Research.

[1]Collaboration discussed before Iteration for better flow.

Virtual (Originally Yokohama, Japan) '21, May 08–09, 2021, Virtual

addresses those needs. However, this needs-first principle is currently sometimes lost in the process when developing AI methods specifically for clinical use. This results in inverted thinking: creating or identifying a novel method, then formulating a suitable clinical problem to which it can be applied. In other words, these methods target "technical novelty" rather than "clinical need".

The issue with starting from the solution instead of the problem is that the ultimate application of these methods to the clinical setting may not be tractable, necessary, or even useful. Any one of these shortcomings can prevent methods from meaningfully reaching the clinic. Currently, many medical AI methods suffer from this disconnect between problems that clinicians find important and those being propelled forward in methods research.

Anecdote. A recent exchange with my clinical collaborator and advisor illustrates this disconnect on a minute scale. As I explained the performance of the computer vision image segmentation method I was developing, he commented on my choice of evaluation metric. While the metric I used is standard in computer vision and offers detailed information about model errors, it would be excessive for our particular clinical use case—the nuanced differences in errors would not be clinically meaningful. The metric of interest in our downstream application would actually be coarser and less complex, instead placing a stronger emphasis on reliability across many image conditions (rather than fine-grained accuracy on any given image). Taking a step back, this disparity between metrics of interest reflects a small gap between the clinical problem of interest and the AI methods objective. Luckily, I can frequently adjust objectives based on regular discussions with clinical collaborators; however, it doesn't take much to imagine how these differentials can compound over the course of AI methods development to an unresolvable state at project's end (especially without collaborations like these in place, which we return to soon).

To be clear, this is not to discourage methods-centered innovation, which has always been crucial for propelling discoveries forward and should proceed unhindered in parallel. Technology-focused methods development can provide useful solutions for bringing AI to the clinic, especially since we often seek to leverage methods that already exist. This principle of Needs is only to say that when aiming to prepare AI methods for healthcare in practice, the choice and development of methods should hinge on true clinical need. [53] calls this important sweet spot between theory and practice Pasteur's Quadrant, which [41] elegantly puts as "the quest for fundamental knowledge within a specific use context".

The ultimate solution to this issue is orienting medical AI researchers toward using a needs-focused approach for formulating medical AI problems to tackle. When developing methods for the clinic, method novelty should fall second to user needs. Just as the success of innovative production-ready products like the iPhone hinged on leveraging and combining existing technologies exceptionally well rather than contriving new inventions—that is, cleverly 'reinventing' over 'inventing' [9, 22]—so too should medical AI methods. (We will return to thinking about medical AI as a product later). Methods should be in service of a genuine clinical need, as experienced in one or more concrete contexts.

Concrete Proposals. In the near term, there are two exciting directions that we in the AI community can take to push forward this way of thinking: **needs-first incentives in publication venues and funding** and **deep collaborations**.

The first is further developing mainstream, reputable, and high-impact venues the likes of *MICCAI*, *MLHC*, and *CHI*, and journals such as *Nature Medicine*, *JAMA*, and *npj Digital Medicine* for researchers to share findings that prioritize medical use cases and implementations rather than only algorithmic novelty. Some journals like the latter are beginning to establish clear author guidelines that underscore this priority [29]. Bolstering these publication venues will also foster interdisciplinary conversations (such as this workshop) and open forums that can align methods with needs. In addition to fostering strong domain-specific venues, another parallel effort should be strengthening the presence and status of needs-focused papers in existing general methods venues. Equalizing so-called "applications papers" and "methods papers" in general publication venues can create larger platforms for these works and build interdisciplinary bridges. In addition to publication incentives, institutional leadership and research funders can also encourage these efforts by allocating positions and funding to needs-first work. Having publication venues, institutional leadership, and funding sources put more systematic incentives in place to encourage AI researchers in academia to consider the practical applications of their work to the clinical setting will be key in closing the gap between methods development and clinical need.

Another key ingredient to foster needs-first research, as discussed in my anecdote, is developing strong, deeply engaged collaborations between medical and technical teams for tackling clinical problems. This naturally leads into the next pillar, Collaboration.

2.2 Collaboration

Collaboration is a fundamental property of design; creation does not happen in a vacuum. Successful design efforts require clear communication and equal understanding among all involved parties. In interdisciplinary domains such as medical AI, it's thus critical to not only establish regular communication between technical and clinical teams (interdisciplinary collaboration), but also *within* these two groups (intradisciplinary collaboration). Information silos in either of these two channels, as present in medical AI, will stunt productivity and quality.

2.2.1 Interdisciplinary Collaboration. Technical and clinical teams have different skill sets and knowledge relevant to medical AI development, which can be synergistic and powerful, but with siloed information, can lead to incomplete, and even divergent understandings of the end product. In the Needs pillar, we saw that information silos of the clinical community can create misaligned objectives within the AI community. The inverse problem also exists; medical teams are not always thoroughly educated on the inner workings of technical teams' developed algorithms, how to use them, nor their intended uses and scope. This siloed knowledge that stays within the technical team results in lower clinician fluency and trust of AI technologies, which naturally leads to lower adoption.

To begin solving this problem, we must examine the incentives at play. Academic research is unlike industry product development, where the incentive structures are often clear for product developers to build products that end users are willing to buy at a certain price point. Instead, institutional incentives for AI researchers have

traditionally been to publish high-impact papers that are cited and used by other researchers in future innovations. Clinicians, on the other hand, most value their patients' experiences and well-being, along with their own workflow efficacy and professional reputation. While these value structures appear divergent, we can motivate regular, deep collaborations by noticing points of mutual benefit. As more and more publication venues recognize the importance of needs-focused research, AI researchers will be keen to develop methods that reach the clinic and aid clinicians; however, we may lack the domain knowledge to make it practical. As clinicians, such as my collaborators, realize the potential of AI to transform the entire clinical experience and streamline their workflow, they will want to spearhead efforts to bring new technologies to the clinic, thereby positively impacting their practice; however, they may not possess the AI toolkit to make it a reality. Establishing effective collaborations between AI methods creators and end users can meet needs on both sides of this equation.

Concrete Proposals. Concretely, this means we as AI researchers need to **actively cultivate integrated cross-disciplinary teams**—to keep clinicians firmly in the medical AI development loop and establish regular information exchange throughout each step of the process, from need-finding to prototyping to deployment.

In my advisors' experience, as well as mine, clinicians are eager to offer their expertise in methods design as well, and their perspective can only make it stronger. Depending on the project and access to collaborators, this may take many forms, such as weekly check-in meetings or less frequent deep-dives into technical methods that motivate model choices, both of which I have done during my clinical collaborations. As clinicians have very busy schedules, AI researchers must be mindful of and efficient with the time we have with our collaborators. The key is maintaining regular exchanges and establishing genuine partnerships that foster information flow to avoid accumulation of knowledge silos. These exchanges additionally give the clinical team opportunities to share feedback and insights with the AI team, as my clinical collaborator readily did in my anecdote, allowing researchers to adapt our methods to better fit clinical needs. For instance, clinicians can offer crucial insights for tradeoffs and priorities: do we care more about accuracy or latency, interpretability or performance, generalization or specialization? These conversations might make use of user interview approaches like Tangible Business Process Modeling [21], which enable technical teams to visually co-design, or collaboratively brainstorm and iterate on model processes in ways that can be readily understood by non-technical domain expert interviewees.

Furthermore, to truly earn buy-in from clinicians, the end users of these methods, it is crucial to underscore that these technologies are built to *aid* clinicians as opposed to *replace* them. Through multiple case studies, [55] emphasizes that medical technologies' greatest power lies in augmenting doctors' abilities so that they can better do their jobs, and warns against the trap of the reverse. Bringing this mindset into practice also requires spending adequate time training clinicians on developed AI methods, including how they work, signs of when they are uncertain or malfunctioning, appropriate usage settings, and importantly, their limitations and intended scope. This will reduce the mystery of, and thus increase the trust in the notoriously "black box-like" methods we create.

2.2.2 Intradisciplinary Collaboration. Information silos are equally obstructive within disciplines. Scientists across many disciplines concur that human intelligence is far-reaching because of our collective intelligence and generational accumulation of knowledge [23]. In the field of AI, the research community has set an excellent precedent of creating open, systematic repositories of information—benchmark datasets such as ImageNet [48] for image classification, GLUE [57] for natural language understanding, RoboNet [12] for robot learning, Common Voice [2] for speech recognition, YouTube VOS [59] for video object segmentation, and COCO [38] for object classification, detection, segmentation, and captioning; strong encouragement of reproducible and documented code in leading conferences like *CVPR* [2], *NeurIPS* [3], and *EMNLP* [4]; and plug-and-play high-level APIs for common methods within frameworks like PyTorch [5], TensorFlow [6], and Keras [7].

However, within specialized sub-disciplines like medical AI, the methods community has room for improvement in creating these centralized knowledge banks, or practicing "open science" [43]. In particular, the sensitive nature of medical data and tighter regulations on release feeds into the weaker cultural expectation in the medical community that data should be released [42]. Thus, many published works do not share their data (and sometimes even code), creating data silos and making reproducibility difficult. A quote from a recent discussion with my clinical collaborators stuck out to me: "With medical AI projects, data is a pain. There is nowhere you can go to get it; you have to build it from scratch yourself." Over time, this lack of widely available, systematically documented and easily usable building blocks in turn can slow down progress of the community as a whole and thus also the development of quality, well-tested algorithms for deployment at point-of-care.

The ideal solution would be for all medical AI researchers to share their datasets publicly and document, benchmark, and publish their code, as well as an academic culture that encourages this. However, historic precedents and case-by-case constraints often render much of this impractical.

Concrete Proposals. Instead, we can look to immediate, actionable opportunities for the medical AI community to improve information flow: taking after the general AI community, creating more **large-scale public benchmark datasets** for systematicity; developing **standard documentation guidelines for data and code**; **releasing documented data and code** wherever possible; and creating **platforms for crowdsourcing clinical annotations**. After motivating these proposals, I will discuss the incentives at play that may hinder them and how we may address these.

As an example, in my home lab at Stanford and beyond, [6, 14, 26, 27] have set the example of releasing large, challenging benchmark datasets for systematic and consistent evaluation of methods, which can aid in selecting methods to push to the clinic. Some medical AI fields also have established data consortiums that curate and maintain databases of interest to a specific community [1, 25]. The field would benefit from further release and curation of datasets

[2] http://cvpr2020.thecvf.com/
[3] https://nips.cc/
[4] https://2020.emnlp.org/
[5] https://pytorch.org/
[6] https://www.tensorflow.org/
[7] https://keras.io/

in specialized medical AI domains to establish high-quality, standard evaluation metrics. Developing standard documentation templates for datasets and models, an effort proposed in the general AI community [19, 40], can add further rigor and consistency among datasets and algorithms. These concrete guidelines for data processing to comply with relevant regulations, and creating streamlined pipelines to do so, can also help to foster a culture of data and code sharing. Even on a smaller scale, clearly documenting and releasing high-quality, modular, bug-free, and readable code and data in repositories such as Zenodo [8], Dryad [9], and GitHub [10] is crucial to allow others to reproduce, learn from, and scale our results; for instance, for my projects, we plan to anonymize and publish our data for posterity, as well as release clean and usable code. Finally, due to the specialized nature of medical AI data, it's currently very difficult to quickly crowd-source annotations from multiple medical experts; I see an opportunity in this area to build secure platforms that match AI scientists and domain experts from different institutions to collectively construct datasets. As a step further, enabling clinicians to create protocols and train certified non-experts to annotate accurately could also widen this bottleneck.

As with interdisciplinary collaborations, we should acknowledge institutional cultures that may disincentivize these proposals. For the general AI community, open-sourcing information is deeply embedded into the innovation workflow; method reproducibility enables future work to build directly on the work (and thus cite it). However, in the medical community, there is much higher risk aversion to sharing information, and understandably so. Not sharing information with other medical institutions often offers a competitive advantage, while sharing doesn't provide any reward. Misaligned incentives between authors and publishers also pose barriers; for instance, open-access publishers earn revenue from article processing charges, but this cost barrier increases barriers for unfunded scientists to publish their work [44]. These cultures exist for a reason, and thus we shouldn't necessarily expect a culture shift. Yet, since preliminary efforts in this domain and plenty of work in other domains have illustrated the benefit of pooling knowledge, it would be wise to further incentivize this initiative in medical AI. In particular, all four of the above efforts can be incentivized with increased recognition and funding. A recent article in the *New England Journal of Medicine* [7] proposed a standardized authorship credit for data collectors and stewards on publications as an incentive for openly sharing their data. Furthermore, a recent workshop report from the *National Academies of Sciences, Engineering, and Medicine* [44] highlighted several possible solutions, including having institutional leadership explicitly endorse open science, establishing regular discourse with leadership and membership about the importance of openly sharing work, having senior faculty set examples by publishing in open-access journals (which can also elevate the status of open access journals in the community), funding positions (e.g. through philanthropic funding) dedicated to open-sourced work, and having funders set data sharing as an expectation at the outset of grant applications and selecting awardees with this in mind. Establishing these efforts in the medical AI community can help to incentivize and enable open science.

[8]https://zenodo.org/
[9]https://datadryad.org/stash
[10]https://github.com/

2.3 Iteration

In addition to need-finding and collaboration, a core piece of design thinking is rapid iteration. As a burgeoning field, medical AI can look to the realms of industry product development and delivery science to see the importance of rapid prototyping. The concept of the MVP, or minimally viable product, is at the heart of launching new products and companies. The key is creating the simplest possible working prototype, then immediately trying it out, collecting feedback, and course-correcting in an iterative fashion. Just like our previous discussion of integrated and regular exchanges with collaborators, it enables faster discovery of paths of greatest value using minimal sunken cost and time. As examined in prior studies that formulate delivery science frameworks for AI in healthcare [28, 36], iteration should occur at every administrative level of methods development, from conceptual to procedural to implementational.

Though rapid iteration is seemingly straightforward in concept, applying this principle to medical AI development isn't trivial. An ideal long-term goal is establishing 'playpen' environments in real clinical centers that allow algorithms to be run on real data without affecting patient care; however, this likely involves lengthy approval times and restrictions. While we computer scientists may be accustomed to rapid development and deployment, the medical field is highly regulated and therefore moves slower and has longer cycles. Data is scarce, testbeds are highly restricted (and for good reason), and clinicians are also understandably busy and many collaborate in their spare time.

Rather than a source of frustration, we should see this as an enormous opportunity to get creative within our constraints. A particularly exciting precedent has been set by AI for autonomous vehicles (AV). Similar to medical settings, AV is in a similar position of needing low-stakes, realistic testing grounds for eventual deployment in high-stakes situations.

Concrete Proposals. AV has crafted two notable solutions: **realistic data synthesis methods** [11, 49, 58, 60] and **high-fidelity, realistic simulated driving environments** [13, 18, 35, 37]. These solutions enable much faster iteration because synthetic data alleviates the bottleneck of data scarcity, while high-fidelity simulation environments create unlimited low-stakes, but realistic testing grounds. Drawing parallels to medical AI, development of realistic simulation and data synthesis methods has the potential to drive methods testing and deployment forward without the risk of deploying immature technologies to the bedside. Recent efforts in dermatology [20], Electronic Medical Records [47], and pathology [34, 45] have demonstrated the great potential of these methods, and it will be exciting to extend this work beyond diagnosis into other aspects of the clinical workflow.

Usually, only MVPs and proofs-of-concept that are thoroughly evaluated, iterated upon, and proven to do well are green-lit for trials and deployments in real clinical settings. Taking the time to do ample iteration to prove method feasibility and reliability opens the gates to potential point-of-care and larger-scale use. Thus, patiently embracing the iterative design process enables the development of methods vetted by technical teams and trusted by clinicians, which can meaningfully bridge the AI-to-clinic translational gap.

In fact, the lessons to take from industry product design run even deeper. A less obvious implication of reframing medical AI as a

product being designed and iterated upon is seeing each algorithm as a solution to the needs of an end user—for instance, a clinician or surgeon—rather than a bundle of code, which is more commonly the state in which many AI projects are left. Creating human-centered AI methods has never been more important, especially as these technologies become more powerful and it becomes easier to forget their ultimate users and usage contexts. As discussed under the Collaboration pillar, clinicians care profoundly about their patients and workflow, and as a result, how the tools they adopt impact this entire clinical experience. This is where having deep collaborations and conversations with medical partners, understanding needs, and rapidly iterating play a crucial role: creating an AI-enabled *experience*, not just an AI-powered *method*. To repurpose the thesis of C. P. Snow's *The Two Cultures*, enabling the technological and human aspects of the clinical experience to see eye-to-eye would accelerate innovation in both worlds [52].

2.4 Error Prevention and Monitoring

Messy, real-world data poses multitudes of challenges to models, and potential resulting errors are critical to catch due to the high-stakes clinical setting. For instance, models may perform worse for patients of less-represented demographics, and even models that perform well in these cases may still not make explainable decisions, or they may degrade over time as the nature of incoming data changes, deviating from the original distribution the model learned from (known as data 'distribution shift' [32]). Thus, to create AI methods that can be safely integrated into clinical workflows, there must be (1) certain guarantees about model bias, explainability, and robustness prior to deployment, and (2) reliable ways (both manual and automatic) to closely monitor and evaluate models over time even after deployment. Both of these mechanisms are currently lacking in medical AI methods development.

The ideal solution to these two complementary axes is, of course, to ensure model creators develop unbiased, explainable, and robust models, and to create and integrate these continuous monitoring systems into the clinical environment. However, each of these challenges in themselves is a whole field of study, and just as discussed under the Iteration pillar, integration of new systems into the clinical environment happens on large timescales. Acknowledging this, I still believe there are things we as AI research scientists can do in the short term to support this long-term vision.

Concrete Proposals. For pre-deployment efforts, there continues to be plenty of exciting **AI research in fairness, transparency, robustness, interpretability, and explainability** that aim to address the first axis of problems [4, 5, 10, 17, 33, 54, 56, 61]. Much of this research has also been applied to and extended into the medical AI domain [15, 16, 30, 31, 46, 50, 51, 62]. It is critical to continue to encourage and **incentivize this research direction in major venues** similar to the way described in the Needs pillar, with both recognition and funding; furthermore, these discussions should be integrated with venues of clinical interest so that domain experts and end users can provide guidance on standards and expectations of systems in these areas.

But as we know well, basic research must translate well to application. To rigorously evaluate models for deployment, the AI and medical communities should jointly develop **standardized compliance metrics** for each of these areas. One key way to do this is, as described in the Intradisciplinary Collaboration sub-pillar, creating **large, public benchmark models and corresponding metrics** that assess particular aspects of model performance and generalization. This could mean curating diverse datasets across patient demographics to test model bias and challenging datasets with unconventional or out-of-distribution cases to test model robustness. For areas such as interpretability and explainability, it is critical to **collaborate with clinicians** to understand the required level of transparency into the inner workings of models and how they want to interface with models. Just as medical AI researchers should place more emphasis on needs-first development, [24] found that within interpretability research efforts for medical AI, there is a mismatch with methods and needs that can be bridged with interdisciplinary communication.

Finally, for post-deployment efforts, **continuous monitoring and improvement of models** is necessary to maintain them at expected levels of performance. Even if these monitoring systems cannot be immediately implemented in the clinical environment, it is important to develop both automatic and human-queried ways to monitor model performance for irregularities. In particular, we should **standardize measures of model confidence and accuracy** and develop **self-reporting mechanisms** for when these metrics slip. These mechanisms should be fully integrated into model pipelines and iterated upon the same way we handle training and evaluation. They should also be easily accessible and readable by human inspectors. In tandem, we should continue developing methods for adapting models quickly to new and differently-distributed data to help mitigate problems like distribution shift. To incentivize these efforts, regulatory institutions, academic and institutional leadership, publication venues, and research funders should underscore the necessity of these measures for medical AI efforts to be considered eligible for deployment. Taken together, these exciting directions will make models more adaptive and monitorable, and therefore better-suited to clinical settings.

3 CONCLUSION

Despite great advances in medical AI methods development, translation of these methods to the clinic has been much less widespread. From a medical AI student researcher's perspective, this paper identifies weaknesses in four key principles (**N**eeds-Focused **I**nnovation, **I**teration, Inter- and Intradisciplinary **C**ollaboration, and **E**rror Prevention and Monitoring) that contribute to the translational gap between AI methods and clinical implementation. It then discusses ideal solutions for these areas in the long term, and puts forth immediate, concrete proposals for us in the AI community that will enable us to help push these solutions forward. While the road ahead is long, the hope of this paper is to spark and synthesize pieces of this conversation that can propel our collective progress toward bridging the AI-to-clinic translational gap.

ACKNOWLEDGMENTS

J. Gong would like to thank Prof. Yeung and Prof. Holsinger for their support and insightful discussions, Dr. Currano and Dr. Sirkin for their helpful guidance and suggestions, as well as Jan Auernhammer, Benjamin Newman, Joseph Makokha, Lawrence Domingo, and Xiao Ge for their thoughts and suggestions.

REFERENCES

[1] 2020. *Innovative Medicines Initiative.* Retrieved February 15, 2021 from https://www.imi.europa.eu/

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 4211–4215.

[3] Jan Auernhammer. 2020. Human-centered AI: The role of Human-centered Design Research in the development of AI. In *Proceedings of Synergy - DRS International Conference 2020*, S. Boess, M. Cheung, and R. Cain (Eds.). Online. https://doi.org/10.21606/drs.2020.282

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences* (2020). https://doi.org/10.1073/pnas.1907375117

[6] Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine* 15, 11 (11 2018), 1–19. https://doi.org/10.1371/journal.pmed.1002699

[7] Barbara E. Bierer, Mercè Crosas, and Heather H. Pierce. 2017. Data Authorship as an Incentive to Data Sharing. *NEJM* 376 (April 2017), 1684–1687. https://doi.org/10.1056/NEJMsb1616595

[8] Stanford Biodesign. [n.d.]. *Biodesign Innovation Process.* Retrieved February 15, 2021 from https://biodesign.stanford.edu/about-us/process.html

[9] Brian X Chen. 2011. *Always on: how the iPhone unlocked the anything-anytime-anywhere future–and locked us in.* Da Capo Press, Incorporated.

[10] Hanjie Chen and Yangfeng Ji. 2020. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4236–4251. https://doi.org/10.18653/v1/2020.emnlp-main.347

[11] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. 2020. Curriculum Model Adaptation with Synthetic and Real Data for Semantic Foggy Scene Understanding. *International Journal of Computer Vision* 128, 5 (May 2020), 1182–1204. https://doi.org/10.1007/s11263-019-01182-4

[12] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. 2019. RoboNet: Large-Scale Multi-Robot Learning. In *CoRL 2019: Volume 100 Proceedings of Machine Learning Research*. arXiv:1910.11215 [cs.RO]

[13] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 78)*, Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (Eds.). PMLR, 1–16. http://proceedings.mlr.press/v78/dosovitskiy17a.html

[14] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542 (Jan. 2017), 115–118. https://doi.org/10.1038/nature21056

[15] Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable Clinical Decision Support from Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1478–1489. https://doi.org/10.18653/v1/2020.emnlp-main.115

[16] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. 2019. Adversarial attacks on medical machine learning. *Science* 363, 6433 (2019), 1287–1289. https://doi.org/10.1126/science.aaw4399

[17] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[18] N. Fouladinejad, N. Fouladinejad, M. K. Abd Jalil, and J. M. Taib. 2011. Modeling virtual driving environment for a driving simulator. In *2011 IEEE International Conference on Control System, Computing and Engineering*. 27–32. https://doi.org/10.1109/ICCSCE.2011.6190490

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. (2018).

[20] Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. 2020. DermGAN: Synthetic Generation of Clinical Skin Images with Pathology. In *Proceedings of the Machine Learning for Health NeurIPS Workshop (Proceedings of Machine Learning Research, Vol. 116)*, Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones (Eds.). PMLR, 155–170. http://proceedings.mlr.press/v116/ghorbani20a.html

[21] Alexander Grosskopf, Jonathan Edelman, and Mathias Weske. 2010. Tangible Business Process Modeling – Methodology and Experiment Design. In *Lecture Notes in Business Information Processing*, S. Rinderle-Ma, S. Sadiq, and F. Leymann (Eds.), Vol. 43. Springer, Berlin, Heidelberg, 489–500. https://doi.org/10.1007/978-3-642-12186-9_46

[22] Lev Grossman. 2007. Invention of the year: The iPhone. *Time Magazine Online* 1 (2007).

[23] Joseph Henrich. 2016. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter.* Princeton University Press. https://psycnet.apa.org/record/2016-18797-000

[24] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 068 (May 2020), 26 pages. https://doi.org/10.1145/3392878

[25] National Human Genome Research Institute. 2020. *The Alliance of Genome Resources.* Retrieved February 15, 2021 from https://www.genome.gov/Funded-Programs-Projects/Computational-Genomics-and-Data-Science-Program/The-Alliance

[26] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *AAAI* (2019).

[27] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 160035 (May 2016). https://doi.org/10.1038/sdata.2016.35

[28] Kenneth Jung, Sehj Kashyap, Anand Avati, Stephanie Harman, Heather Shaw, Ron Li, Margaret Smith, Kenny Shum, Jacob Javitz, Yohan Vetteth, Tina Seto, Steven C Bagley, and Nigam H Shah. 2020. A framework for making predictive models useful in practice. *Journal of the American Medical Informatics Association* (12 2020). https://doi.org/10.1093/jamia/ocaa318 ocaa318.

[29] Sujay Kakarmath, Andre Esteva, Rima Arnaout, Hugh Harvey, Santosh Kumar, Evan Muse, Feng Dong, Leia Wedlund, and Joseph Kvedar. 2020. Best practices for authors of healthcare-related artificial intelligence manuscripts. *npj Digital Medicine* 3, 134 (Oct. 2020). https://doi.org/10.1038/s41746-020-00336-w

[30] Amit Kaushal, Russ Altman, and Curt Langlotz. 2020. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA* 324, 12 (Sept. 2020), 1212–1213. https://doi.org/10.1001/jama.2020.12067

[31] Newton M. Kinyanjui, Timothy Odonga, C. Cintas, Noel C. F. Codella, R. Panda, P. Sattigeri, and Kush R. Varshney. 2019. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. In *NeurIPS 2019 Workshop on Fair ML for Health*.

[32] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. WILDS: A Benchmark of in-the-Wild Distribution Shifts. (2020).

[33] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GeDi: Generative Discriminator Guided Sequence Generation. (2020).

[34] Adrian B. Levine, Jason Peng, David Farnell, Mitchell Nursey, Yiping Wang, Julia R. Naso, Hezhen Ren, Hossein Farahani, Colin Chen, Derek Chiu, Aline Talhouk, Brandon Sheffield, Maziar Riazy, Philip P. Ip, Carlos Parra-Herran, Anne Mills, Naveena Singh, Basile Tessier-Cloutier, Taylor Salisbury, Jonathan Lee, Tim Salcudean, Steven J.M. Jones, David G. Huntsman, C. Blake Gilks, Stephen Yip, and Ali Bashashati. 2020. Synthesis of diagnostic quality cancer pathology images. *bioRxiv* (2020). https://doi.org/10.1101/2020.02.24.963553

[35] Kunming Li, Yu Li, Shaodi You, and Nick Barnes. 2017. Photo-Realistic Simulation of Road Scene for Data-Driven Methods in Bad Weather. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 491–500. https://doi.org/10.1109/ICCVW.2017.65

[36] Ron C. Li, Steven M. Asch, and Nigam H. Shah. 2020. Developing a delivery science for artificial intelligence in healthcare. *npj Digital Medicine* 3, 107 (Aug. 2020). https://doi.org/10.1038/s41746-020-00318-y

[37] Wei Li, Chengwei Pan, Rong Zhang, Jiaping Ren, Yuexin Ma, Jin Fang, Feilong Yan, Qichuan Geng, Xinyu Huang, Huajun Gong, Weiwei Xu, Guoping Wang, Dinesh Manocha, and Ruigang Yang. 2019. AADS: Augmented autonomous driving simulation using data-driven algorithms. *Science Robotics* 4, 28 (2019). https://doi.org/10.1126/scirobotics.aaw0863

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing,

Cham, 740–755.
[39] Robert H. McKim. 1980. *Experiences in Visual Thinking*. Brooks/Cole Publishing Company.
[40] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. *FAT** (2018).
[41] Donald A. Norman. 2010. The research-practice gap: the need for translational developers. *Interactions* 17, 4 (July 2010). https://doi.org/10.1145/1806491.1806494
[42] Institute of Medicine. 2013. *Sharing Clinical Research Data: Workshop Summary*. The National Academies Press, Washington, DC. https://doi.org/10.17226/18267
[43] National Academies of Sciences, Engineering, and Medicine. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25116
[44] National Academies of Sciences, Engineering, and Medicine. 2020. *Advancing Open Science Practices: Stakeholder Perspectives on Incentives and Disincentives: Proceedings of a Workshop–in Brief*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25725
[45] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. 2020. PathologyGAN: Learning deep representations of cancer tissue. In *Medical Imaging with Deep Learning*.
[46] Jean-Francois Rajotte, Sumit Mukherjee, Caleb Robinson, Anthony Ortiz, Christopher West, Juan Lavista Ferres, and Raymond T Ng. 2021. Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary. (Jan. 2021).
[47] Anat Reiner Benaim, Ronit Almog, Yuri Gorelik, Irit Hochberg, Laila Nassar, Tanya Mashiach, Mogher Khamaisi, Yael Lurie, Zaher S Azzam, Johad Khoury, Daniel Kurnik, and Rafael Beyar. 2020. Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med Inform* 8, 2 (20 Feb 2020), e16492. https://doi.org/10.2196/16492
[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y
[49] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. 2018. Model Adaptation with Synthetic and Real Data for Semantic Dense Foggy Scene Understanding. In *European Conference on Computer Vision (ECCV)*. 707–724.
[50] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. https://doi.org/10.1145/3351095.3372827
[51] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2019. Fair Predictors under Distribution Shift. In *NeurIPS 2019 Workshop on Fair ML for Health*.
[52] Charles P. Snow. 1993. *The Two Cultures*. Cambridge University Press.
[53] Donald E. Stokes. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution Press.
[54] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 107–118. https://doi.org/10.18653/v1/2020.emnlp-demos.15
[55] Robert Wachter. 2015. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age*. McGraw-Hill Professional. https://books.google.com/books?id=qO-VBgAAQBAJ
[56] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. 2019. Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[57] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In the Proceedings of ICLR.
[58] Mingyun Wen, Jisun Park, and Kyungeun Cho. 2020. A scenario generation pipeline for autonomous vehicle simulators. *Human-centric Computing and Information Sciences* 10, 24 (June 2020). https://doi.org/10.1186/s13673-020-00231-z
[59] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. 2018. YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark. *ECCV* (2018).
[60] Zhenpei Yang, Yuning Chai, Dragomir Anguelov, Yin Zhou, Pei Sun, Dumitru Erhan, Sean Rafferty, and Henrik Kretzschmar. 2020. SurfelGAN: Synthesizing Realistic Sensor Data for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[61] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. 2019. Towards Interpretable Face Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
[62] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.