**CRAIG:** Hi, I’m Craig Smith and this is Eye on AI.

This week, I speak to Daniel Ho, a Stanford professor and associate director of Stanford’s Institute for human centered artificial intelligence, about the proposed National Research Cloud, also known as the National AI Research Resource – an effort to expand the data and compute available to academic researchers. The initiative was included in the final report by the National Security Commission on AI.

If all of these names confuse you, the bottom line is that academic researchers have trouble getting access to datasets and compute on the scale available to researchers in private companies like Google and Amazon. People like Daniel are focused on correcting that.

**CRAIG:**

**CRAIG:** Hi, Dan, why don't we start by having you introduce yourself a little bit of your background, what you do at HAI or H.A.I.

**DANIEL:** Great to be here. My name is Dan Ho. I'm a professor at Stanford where I teach both at the law school and in the political science department. And I'm also an associate director at Stanford's Institute for Human-Centered Artificial Intelligence or Stanford HAI and I also direct the Regulation, Evaluation, and Governance Lab, or the RegLab, which thinks a lot about public sector technology.

**DANIEL:** In late 2019, the two co-directors of Stanford HAI Fei-Fei Li and John Etchemendy issued a call for what we have been calling the National Research Cloud, but which in the legislation is referred to as the National AI Research Resource or the NARR but I'll just use NRC.

**DANIEL:** And there was a tremendous amount of enthusiasm. There were 21 other top computer science peer institutions that issued a call for this NRC. The National Security Commission for AI endorsed the idea and it ended up in legislation early this year in January to create a task force to think about how to design the National Research Cloud, to provide a resource and ease the bottleneck for AI research, compute and data.

**DANIEL:** And the way that our team got involved is that it's a great idea, but there are lots of interesting challenges that emerge in thinking about the implementation of such a National Research Cloud. And so, we convened a policy practicum here at Stanford, where we had over a dozen business, law, and engineering students join a group of us, really to think about the various, crucial questions of how to design and implement such a National Research Cloud. And the co-instructors with me on that policy practicum that are also on the report are Russell Wald who is the policy director at HAI, Jen King who is the Privacy Fellow at HAI, and Chris Wang.

**CRAIG:** And the recommendations then go to the task force. Was Stanford asked to make these recommendations or is it because you guys felt that it's important to put your thoughts forward?

**DANIEL:** We were not formally asked by the task force, but really thought that we were well-situated given both Stanford's amazing engineering, law and public policy-oriented students to try to shed some light on the major questions that the task force would have to grapple with.

**DANIEL:** I think the most important takeaway that really animates both the legislation and our report is that the NRC is an opportunity to rebalance the AI innovation ecosystem to foster non-commercial and fundamental research given how much of the shift over the past few years has been towards a private sector or applications.

**CRAIG:** Because the compute and the data is really concentrated at this point in the private sector, in the Facebook and Amazon and Microsoft's of the world. And this would democratize. That was the concept, to democratize. Both compute and data.

**DANIEL:** That is very much the goal. There are cutting edge advances that depend critically on large scale compute and data. And the NRC is really an attempt to rebalance this, to provide more equitable access so that it's not only the sort of top technology players, elite academic institutions that have access to this kind of a resource, but really to enable a much wider range of voices to be able to contribute. I think one statistic here that is illuminating is that in 2011, about 50% of AI PhD graduates were equally likely to go into the academy or into industry.

**DANIEL:** And 10 years later, about two thirds go into industry. And compute and data are a really big component to that. I'm reminded of one of my PhD students who delayed her return to campus, primarily to be able to retain access to the computing resources of a large company. And that of course really does have pretty profound effects on the kinds of research questions that are being asked.

**DANIEL:** One technologist famously said the best minds of our generation are thinking about how to make people click ads. And I think part of the goal here is actually to have a little bit of rebalancing to enable the kind of non-commercial long-term fundamental research that is more at home within academia.

**CRAIG:** Reverse that brain drain and effect, and not only balance human resources, but the research itself because academic institutions are less likely to focus on relatively trivial applications like ad serving. Are there other organizations making recommendations to the task force that you're aware of and what happens once the task force sorts through the recommendations or makes its own recommendations? Does that then go to various committees in Congress that have to come up with legislation? Is there likely to be a single piece of legislation or is this going to end up broken into bits and embedded in different kinds?

**DANIEL:** There was a request for information for all interested parties to submit their perspectives on the design of the National Research Cloud, which closed, I believe October 1st of this year. So, we submitted part of this report as part of that request for information. The task force has a series of public meetings to consider various elements, and these are open for anyone to attend. And ultimately the task force has to issue both an interim report and a final report making its recommendations as to how the National Research Cloud should be designed. And really, depending on those recommendations there may be legislative action to enact something like this, but it is also possible to think about some elements of this that may not necessarily require a direct legislative intervention. So, it really depends on what the ultimate recommendations are. For instance, one of the questions we considered in the report is should the National Research Cloud be structured essentially as a new government agency.

**DANIEL:** And obviously if we're talking about the creation of a new government agency, the technical infrastructure and the very challenging problem of providing public sector data, that would require enabling legislation. But what the next steps are really will hinge on what the recommendations of the task force might be.

**CRAIG:** On that particular question, you guys recommended that the NRC instituted as a federally funded research and development center, and then eventually as a public private partnership. I'm not very familiar with FFRDCs so maybe could you just take a minute and explain what an FFRDC is, how it becomes instituted?

**CRAIG:** Whether that requires legislative action?

**DANIEL:** So, when we thought about the institutional form that the NRC should take, we really have to think about the combination of both compute and data access. There are really clear statutory prohibitions on interagency data sharing, most notably in the Privacy Act of 1974. And so that led us really to think about vehicles like the federally funded research and development centers, which serve a particular agency, and as a result would have more streamlined mechanisms by which to get data access.

**DANIEL:** The central insight of the National Research Cloud is that to foster a healthier AI research ecosystem, it's not very effective, to focus exclusively on the compute side or exclusively on the data side.

**DANIEL:** tremendous complementarity between the two.

**CRAIG:** But how do you create an FFRDC.

**DANIEL:** Federally funded research and development centers are created by government agencies by contract. And there's a history of FFRDCs being created with particular research institutions. The RAND Corporation, or the MITRE Corporation, or I think actually even Stanford's Linear Accelerator Center can be structured as such, and so part of what we were really grappling with in the report is how to stand up something like this without having to undergo the really significant overhead of creating a completely new institution.

**DANIEL:** That was what we saw as a potential benefit of the FFRDC. Private public partnerships, which you also mentioned. One of the most interesting examples of this that we got in recent years was actually a private public partnership, involving the NSF-funded XSEDE initiative for high performance compute resources for COVID-19 research.

**DANIEL:** There was a range of industry players, government agencies, and academic institutions that came together to provide that kind of a computing resource. That's one of the potential models for how to think about the private public partnership — that they can be designed in a really wide range of ways,

**CRAIG:** But the bottom line is it would not require the creation of a new agency, it would fit within existing pathways or whatever. The compute side of the NRC. You talk about the choice between building and buying, and it sounds as though you recommend an initial step of buying.

**CRAIG:** In other words, contracting compute resources available with existing cloud providers but eventually building resources within the NRC or by the NRC. Can you talk about that a little bit?

**DANIEL:** Sure. One of the core questions for how to build this computing infrastructure is what we might refer to as the make or buy decision. That is, should the government make and publicly own computing infrastructure or should it go to existing commercial cloud providers to build out this kind of capacity. And the short answer in our report is that we need a hybrid approach for two reasons. One is in the short run, it's going to be very hard to stand up the NRC at the contemplated scale, without some degree of reliance on the commercial cloud, simply because the contracting process to build a new high performance computing facility will take quite a bit of time.

**DANIEL:** Second. We spent some time looking into the way that other national research clouds have thought about this design decision. A national research cloud like Compute Canada, which provides access for academics in Canada, itself came to the conclusion that it's orders of magnitude more cost-effective to build and publicly own infrastructure rather than relying exclusively on the commercial cloud. So, Compute Canada at that point, when it did its calculation said, it's going to be four to 10 times more expensive. If we built something like this using purely commercially available resources when there is relative continuity of compute demand.

**DANIEL:** And there are really great examples of that within the US context. The National Science Foundation has invested in something that is known as the XSEDE network of high-performance computing centers across universities. The Department of Energy built what for several years was the world's fastest supercomputer.

**DANIEL:** The key question there will be, will it be possible to scale up those forms of public infrastructure and make them usable in a way that a lot of researchers have found existing commercial cloud services to be easy to onboard. On the commercial thought side, there's actually a very interesting program, also administered by the National Science Foundation, called the CloudBank program.

**DANIEL:** And the CloudBank program essentially provides subsidized commercial cloud credits for NSF grantees to a range of different commercial cloud services. So, one possible avenue to go through is an NSF type mechanism to expand access to cloud services currently. But the bottom line for us was that given these pretty significant trade-offs, we should really be taking a hybrid investment approach and then learning what seems to be working most effectively.

**CRAIG:** Yeah. And I can see with the hybrid approach, cloud services are continually updating their infrastructure with new chips or accelerators and that sort of thing. Then that's one of the decisions that the private sector has to make all the time, whether the build or buy and people are increasingly moving to the commercial clouds because the development of new computing technology is fast enough that if you build, you're very quickly working with outdated equipment, and if you're in the cloud, you’re ensured to be working on the latest chips and accelerators and things like that.

**DANIEL:** The larger issue that you're raising is a very important one in terms of the incentive structure for commercial cloud providers to continuously upgrade their equipment.

**DANIEL:** I think one of the critiques that one sees of existing government-owned supercomputing infrastructure is that they’re built under contract at one point of time, then they are run for a fixed period. And as a result, you don't see the kind of rejuvenation with new chip sets that might be particularly optimized for deep learning approaches quite in the same way.

**DANIEL:** Of course, that's in part because of the level of investment in public infrastructure. Just as right now, when you start up in AWS instance, you can choose from a range of different GPUs, you can go with an A100 or V100. You can imagine being able to have a system that's built out that allows you to choose between public infrastructure — a server that was built out two years ago or one that was more recently stood up — in the same fashion as you would currently with commercial cloud providers. But that of course is a new way of thinking about accessing certain forms of public infrastructure. And that's why we recommend piloting it and seeing how usable we can make a system like that.

**DANIEL:** Currently, something like the Summit System, which was built between 2018 and 2020, the world's fastest supercomputer, is really only accessible to employees at the National Research Laboratory in which it was housed.

**CRAIG:** I talked to Andrew Moore about the data side, and the way he envisioned it is it would be a kind of one-stop shop where you go into an interface, and can see all the data sets available and then, depending on your level of access, can access those data sets directly. But on the data management side, those data sets would not necessarily reside in a single place.

**CRAIG:** It would be a network of data sets that would be tied together. Is that how you guys envision that?

**DANIEL:** We started off with this report, thinking that the compute provision was going to be the biggest challenge, but actually we spent far more time wrestling with the data access portion of this.

**DANIEL:** And so our report is based on a close reading of the existing legal and policy framework for public. Most importantly, the Privacy Act of 1974, which was enacted in the aftermath of Watergate, where there were really serious considerations of government surveillance of citizens. The Privacy Act of 1974 structurally mandates decentralized data across government agencies. The government agencies are expressly prohibited from having a linkage of different administrative data sets because of the fears of what that would entail. That makes a simple one-stop shop challenging because the Privacy Act does not contemplate that a single agency would be able to link all of these different administrative data sets.

**DANIEL:** It's in part because of that legal landscape that being able to share public sector data sets has been such a profound challenge for modern government. One of my colleagues called it a game of bureaucratic hot potato in terms of the amount of contracting that it takes. Another official we interviewed said it was the single hardest thing he did in his entire public career, trying to think about data access.

**DANIEL:** The National Research Cloud is an opportunity to think about how do we do this in a way that is both secure, privacy-preserving and consistent with the Privacy Act of 1974. And the way we think about that is less as a one-stop shop, where a researcher could click and just access a range of different data sets.

**DANIEL:** It really should be a form of tiered access. Where there are certain government data sets, like oceanic data, that are low risk, that anyone should be able to access, but that for medium and high-risk data sets there be more of a process for researchers to apply and for there to be some approval, subject to greater security constraints, to be able to compute on that data.

**CRAIG:** But nonetheless, there would be, I'm imagining, an NRC portal that people would go through. There would be some permissioning in their login, but depending on their status, they would be able to access directly as you say, this low-level data, and then higher tier, there would be an application and review process. But it's not as though it's going to be a bunch of hyperlinks and you have to chase down these hyperlinks and deal with 20 different organizations.

**DANIEL:** That is one of the challenges of the current way of accessing government data. Yeah. We're very much recommending that the NRC serve as a kind of data intermediary that streamlines data access where currently it's so balkanized that one interviewee said, it's really just a matter of which relationships you have, whether you can get access to a certain government data set. And so in that sense, our proposal is most closely aligned with all the thinking that has been done under the Evidence-Based Policymaking Act, where there has been this call for a national secure data service.

**DANIEL:** I think that the data component really very much aligns with that, and maybe the thing I would add here is what the potential significant benefits are of doing this. I think one example that we point to is the way in which the government made available Landsat satellite imagery starting in the 2000s.

**DANIEL:** It used to be the case that you paid $600 for an image tile of publicly collected data, and one estimate is that by making that data available for free for researchers to be able to assess the state of global warming, habitat modification, economic development led to about three to $4 billion benefits annually. What’s really interesting is how much of that has been about the synergy between compute, because we're talking about really large scale satellite data that can really only be interpreted and analyzed when put on a computing facility that makes it easy to tile that imagery, to bring in different remote sensing techniques, to apply computer vision techniques.

**DANIEL:** And that's an example of the kind of value proposition if we can make it easier to get access to that kind of data and would rebalance our current focus. Right now, it is often easier for researchers to work with certain private kinds of data sets. And that of course, really shifts what kinds of problems AI researchers tend to tackle.

**CRAIG:** Yeah. Andrew Moore made a comment that you have some of the best minds in the world working on toy problems because they just don't have access to the large data sets. On the privacy preserving side, do you get into recommending particular approaches?

**CRAIG:** Different kinds of technologies?

**DANIEL:** Yeah, we spend a chapter thinking about technical measures of privacy and how they relate to privacy preserving access to government.

**DANIEL:** We have a few things that we suggest in that chapter. The first is that we should not think about technical privacy measures as a kind of substitute for access policies. The first and most important thing is to think rigorously about access policies when we start to think about medium or high-risk data. But of course, there are going to be many instances where agencies will be more comfortable releasing certain data sets if they're subject to certain privacy measures.

**DANIEL:** We discuss various approaches in terms of homomorphic encryption, differential privacy, federated learning. And we think there's definitely a place for those kinds of approaches, but it's also worth noting that there's not likely to be a one size fits all solution.

**DANIEL:** Let me give you just one example — differential privacy, which is essentially adding noise to make re-identification impossible, have been shown in certain settings to actually make it harder to assess disparities across demographic groups, because minority groups by nature will have less representation in that data.

**DANIEL:** So, adding noise is going to have some differential impact. And so, to the extent that one of the goals here is for researchers to be able to interrogate AI systems for things like bias, we may, in some instance, want differential privacy as the approach. But other instances, we may want to think about alternatives as well.

**DANIEL:** The other thing that I'll just note here is that the other data access model that, for instance, the US census has for US census micro data, is a kind of data safe room. The idea is you apply with a particular research project. You have a background check, and then you go into a physical room where you're walled off.

**DANIEL:** And you're able to compute on high-risk data, but you're only able to take out stuff that is more aggregated. And one of the things we do recommend in the report is that we need to think, as some agencies have, of the virtual analog to that.

**DANIEL:** It's going to be very challenging if some of these data sets remain in this physical enclave and it would be quite valuable to think about what a virtual data safe room would look like, for instance, key stroke logging, or various other forms of monitoring, but still make it possible not to have to physically take a trip to a small number of these federal statistical research data centers to be able to use that kind of data.

**DANIEL:** There was this amazing effort led by Gary King and Nate Persily to create an organization called Social Science One to partner, in that instance, with Facebook to try to get Facebook data access for academic researchers to study, for instance, what the evidence was of election interference in 2016.

**DANIEL:** But early on in that project, there was this possibility to have a data safe room. And the other was to go with the differential privacy approach. And I think right now the sense is that differential privacy made it much harder for there to be a large release of data sets as was originally contemplated. And that has been one of the real pain points of that initiative. And so, if we had to make that choice over again, would it actually be better to have a data safe room approach given how limited the nature of academic inquiry has been for that project?

**CRAIG:** So what happens from here? What is the timeline? The task force has to produce a final report next year, and then it's anybody's guess. It goes to committees. I would guess there may be some executive action that could happen or some agency actions that could happen right away. Do you have a sense for the timeline and even if it's a hope and a prayer when researchers will be able to access a National Research Cloud.

**DANIEL:** It's hard to make a direct prediction as to what exactly will transpire once the task force issues its report next year. But one of the approaches of our report was to articulate a range of both short-term and long-term solutions.

**DANIEL:** What can be done pretty quickly to get an initiative like this off the ground and what are the longer-term investments. For instance, that very much informs our recommendations about compute. In order to improve access to compute for academics, expanding something like the NSF CloudBank initiative is a really great idea.

**DANIEL:** And then the longer-term investments are about the public infrastructure. The hope would be, that the task force issues a report that is informed by some of our thinking on the topic and that then we would see both potentially executive and legislative action to carry this into implementation.

**CRAIG:** Yeah. There's certainly a sense of urgency in some quarters. That's why there is a task force to get this thing stood up. Again, not to hold you to any prediction, but do you think it's going to take a couple of years to get something started or is it more like five years before researchers will be able to access this kind of institution?

**DANIEL:** I think it will be sequenced. And I very much hope that the recommendations will provide for a kind of sequencing to start with the easy wins first and have those happen in relatively short order. When we're talking about AI research, there are fields where a paper that was published two years ago is referred to as really old.

**DANIEL:** I do think there is urgency to enable non-commercial and basic AI research. I do think there are short term things like expanding the NSF CloudBank program, like prioritizing low-risk data sets that currently sit in many government agencies, but that simply haven't been available to researchers.

**DANIEL:** I would hope that is something that we could start to focus on in the short-term. And then in the longer term, obviously, as you develop the data acquisition strategy for the NRC to turn to more of the sort of medium and high-risk data sets and contemplate proposals, like the virtual data safe room. But when we're talking about lower risk data sets, for instance, when it comes to our understanding of the American justice system, the administrative office of the U.S. courts has what is a pretty inaccessible system about the federal courts and being able to open that up, make that available to researchers would lead to all sorts of interesting applications.

**DANIEL:** And understanding of the way that the US justice system works, or think again of the sort of Landsat example of satellite imagery and what that has done for understanding of the earth.

**CRAIG:** Do you have any sense of how much data currently exists in government silos that could be unlocked for academic research, exabyte levels or some metric.

**DANIEL:** I don't have an estimate of the scale of government data sets. It is large. I would be hesitant to just think about the amount of data. It's really, is a data about something that is socially meaningful. So there's no question that, if every click and interaction on an online platform is being recorded, that is quickly going to amount to very large scales.

**DANIEL:** But what government agencies have is data on some of the most meaningful parts of life, labor, birth, death, and being able to get access to those kinds of data sets would empower AI researchers to tackle some of the most important social problems of our time.

**DANIEL:** We're talking about data from the U.S. census, from the Department of Labor on the job market, from the administrative office of the U.S. courts. If we made access to that kind of data easier, we may have more of the amazing talent that exists think less about advertising and a little bit more about some of the really socially important challenges that are embedded within those kinds of data sets.

**CRAIG:** Yeah. And then the other question on the hardware in the cloud, did you guys consider a specific hardware at all?

**DANIEL:** We did not think about specific hardware but what we did think about a fair bit in terms of the design of computing access, is this flexibility to simultaneously compute on different platforms.

**DANIEL:** Ideally you would have a kind of portability so that it's possible to do cloud bursting so you could imagine a base level access, but then you actually need to be able to run a very large job. Ideally you would be able to do so in a way that is portable across different cloud providers. When you're going from a GPU, that's on AWS toward a TPU based compute on GCP, the software stack may not match. And so that is one of the really interesting pain points — transitioning between different cloud providers.

**CRAIG:** Yeah. You're talking about an orchestration layer in effect. There's a company called Rescale that I don't know if you ran across, but they do just that — they have the relationships with all the different cloud providers, the people using the compute resources, don't have to go to every cloud provider. They go to one company that has this layer that can divide the problem up come up with hybrid cloud solutions and that sort of thing.

**DANIEL:** One of the biggest debates, obviously right now is how do we ensure the ethical development of AI products?

**DANIEL:** And we really make three recommendations in the report. Of how to ensure that the NRC is used in ethical ways. And we're mindful they are, of the fact that this is still very much an evolving field, we have hundreds of different ethics guidelines and they are admittedly still at a pretty abstract level.

**DANIEL:** One of the things we were very cognizant of is the concerns about direct government review, for instance, under vague standards of certain research projects. It's not hard to imagine that one could have a political appointee who has one view on some scientific issue, and that could pose real potential speech, academic speech challenges.

**DANIEL:** Cognizant of that, we make three recommendations in the report of how the NRC can really foster ethical use of the resource. The first is that for researchers applying for beyond base level access or data sets that are higher risk since there will be some application process involved, we follow a really great demonstration project that was done here at Stanford, where through a kind of grant making process, researchers were required to provide ethics impact statements.

**DANIEL:** And we think that's a great idea. Second, we're imagining that academic researchers will get base-level access without having to go through a custom, grant application process because the overhead could be very significant for that. And for that, we think that the NRC should have a process to handle when there are complaints that arise about unethical use of the resource, but we caution also that the standard should be fairly high.

**DANIEL:** The concerns about government review of academic speech and that it should be done by independent body, much in the same way that you have independent merit panels that NIH or NSF will have to review societal broader impact statements.

**DANIEL:** The third thing is that how we best foster ethical reflection in AI development is still very much an ongoing enterprise that we as a community are thinking about. And there are definitely critiques about the reliance on ex ante ethics impact statements. It may be really hard to anticipate the kinds of ethical choices that need to be made in the development process rather than at the very outset of a research project.

**DANIEL:** And there, we really think that one of the potential interesting models comes actually from the way that the biomedical sciences and the NIH in particular have thought about ethics, which is that they have provided funding mechanisms to help bioethicists be embedded within research teams and there, the idea is that as you're making ongoing discretionary calls, it just becomes much more of a conversation to figure out what ethical principles are implicated, how to trade off different values and to build that right into the design process.

**DANIEL:** And so that is something that I think more generally for the field would be valuable.

**DANIEL:** That kind of collaborative, consultative model has a lot of promise. One way of thinking about it is that if you have a purely technical measures of performance, we often ignore how AI systems are likely to be embedded within human institutions.

**DANIEL:** And so having a wider range of perspectives like that embedded in the development process can actually be quite valuable. Rather than thinking about those entirely at the outset, when you're proposing a research project or at the tail end, when a research project has concluded.

**CRAIG:** Yeah. And that would include the legal implications, wouldn't it?

**DANIEL:** Potentially. Yes. One of the reasons why we propose here that the NRC start with faculty at universities is both that it is consistent with the way that federal grants generally work. But also that we at least have a little bit of an ecosystem where researchers are trained on various research compliance matters, and that includes thinking about legal compliance.

**CRAIG:** That’s it for this week’s podcast. I want to thank Daniel for his time. If you want a transcript of this episode, you can find one on our website, eye-on.ai. I’ll be following the topic as it winds its way toward reality.

Remember, the Singularity may not be near, but AI is about to change your world. So, pay attention.