Interfolio, Inc.
1400 K St NW 11th Floor
Washington, DC 20005
p: (877) 997-8807
f: 267-295-8740
e: info@interfolio.com

Document #CB9F80EE5B

Application Credentials
Andrew Ilyas
Jan 3, 2024

At the request of Andrew Ilyas, the following document has been uploaded by Interfolio. The document was not uploaded by the evaluator. Any application or evaluation data provided was entered only because it was required during the upload process. For more information, please contact the applicant or Interfolio at help@interfolio.com.

Recommendation from Prof. Aleksander Madry	Confidential Letter of Recommendation or Evaluation	CONFIDENTIAL
Uploaded January 3, 2024	Uploaded By John Calceta, Interfolio, Inc.	3 pages

FERPA Compliance

This document is confidential which means that the applicant waived their rights to access this letter. Interfolio received this letter directly from the letter writer and delivered it to you without the applicant's intervention. As required by the Family Educational Rights and Privacy Act of 1974, these credentials are transmitted to you on the condition that you not allow other parties access to them without written consent of the applicant.

Questions or Concerns

If you have any questions regarding this web submission, or Interfolio, please contact us at help@interfolio.com, (877) 997-8807 or find us on the web at www.interfolio.com.

Aleksander Mądry
Cadence Design Systems Professor of Computing

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
people.csail.mit.edu/madry

November 5, 2023

To Whom It May Concern:

It is an *absolute privilege* to recommend Andrew Ilyas for a faculty position at your institution. Let me say at the outset: I have had the pleasure of closely working with him (on almost daily basis) for the last six year—first as his undergraduate and, now, as his graduate research advisor and *Andrew is simply phenomenal and off-the-charts.*

Indeed, being a faculty at MIT (and sitting on the faculty search committees) gives me an opportunity to interact with many truly terrific students. Still, *during my more than 11 years of being a faculty, I have not met a better student than Andrew!* And I mean here not just "a better student of mine," but "a better student," period. The mix of sheer talent and brilliance, incredible drive and the unusually mature and focused vision, makes him an *absolute gem and a star.* He is *bound to change the field of machine learning.* In fact, his work has accomplished much of that already by changing our thinking about some of the most core questions around machine learning robustness and reliability, as well as around science of deep learning.

Now, I initially wanted to provide you with a detailed overview of what Andrew has accomplished so far. But then I realized that there is a "problem:" Andrew is so exceptionally productive that there is simply too much to cover. Indeed, his track record of *twenty* NIPS/ICML/ICLR papers (on top of papers in other venues), as well as nearly 10,000 citations, is not only extremely prolific but also truly unusual and impressive for someone who is still in graduate school.

So, let me convey to you the caliber of Andrew's work differently: in *my* career in machine learning, there are three results that I am particularly proud of and view as my key contributions to the field. Andrew played an absolutely instrumental role in establishing—and, in particular, was the first author on—*two* of these three. (And I have a feeling he would have also played such role in the third case, if he was working with me at that time already.)

Let me now describe these two results.

The first one of them revolves around the phenomenon of adversarial examples. Specifically, the fact that ML models tend to fail on inputs that have been perturbed in an adversarially chosen—but imperceptible to humans—manner. This phenomenon is a serious concern for safety and security reasons (and Andrew has a number of pioneering work in this context too, some of them being undergrads-only papers). But, there was also a perplexing mystery here: why these adversarial examples exist in the first place—or, rather, why are they so widespread?

Unravelling this mystery took a while but its resolution—driven in a decisive way by Andrew—was: our real-world adversarially robust models might have to *necessarily* rely on a different type of feature representations than the non-robust ones. In particular, the "robust"

representations—which is what we humans use exclusively when we tackle image recognition tasks—might be unable to utilize all the *predictive* correlations that the dataset supports. This is so since the features capturing these correlations might be "non-robust" (i.e., possible to flip using an imperceptible perturbation) thus making the model vulnerable to the corresponding adversarial example attacks.

Indeed, as our NeurIPS 2019 paper (on which Andrew is the very much deserved first author) demonstrates, the vulnerability of our models to adversarial examples is so widespread *exactly* because these models find the reliance on such "non-robust" features be most helpful for maximizing the test accuracy (which is the only objective they care about and pursue).

The resulting perspective completely transformed the way I (and, from what I was told and could observe, the rest of the field) think about adversarial robustness. In particular, it made it clear that at the core, adversarial robustness is not merely about fixing some "glitches" in our learning method but rather about enforcing a "human prior" on the way our trained model make their decisions. In fact, as we fleshed out in follow-up works, the representations synthesized by robust models end up being significantly more interpretable and easier to manipulate. They also enable us to solve a wide range of image synthesis tasks in a very natural way. (I encourage you to take a look at `http://gradientscience.org/adv/` and the subsequent blog posts.)

Now, the other result I want to describe to you revolves around a fundamental challenge in deep learning: how do we get some confidence in the prediction that deep learning models make? This question attracted a lot of attention of the community—in particular, there is a whole area of ML explainability that focuses on it. However, despite a lot of progress, existing solutions were all lacking in an important way: they are not able to provide real *counterfactual* guarantees.

Andrew's recent work (where he, again, was very much deservingly the first author) has finally been able to change that state of affairs and offer a new quality in this context. Specifically, his ICML 2022 work has put forth a new framework: *datamodels*. This framework can be viewed as providing the first reliable (and scalable) operationalization of the influence functions known from statistics in large-scale machine learning context. And, as a result, it enables one to understand large training datasets in a model-modulated way and, beyond providing a new type of counterfactual-based explanations, gives rise to a toolkit for performing a wide range of other data analysis tasks.

Since this work has came out, my group has been busy developing a variety of ways in which one can build further on this framework. This includes, for instance, new approaches to tackle distribution shift, understand the representation/feature formation in transfer learning and provide principled insight into how different learning algorithm choices impact the data biases that the resulting model has.

We (and, again, Andrew played a key role here) also found a way to majorly speed up and streamline the datamodeling approach further, with the development of TRAK (which appeared as an oral presentation at ICML 2023).

All in all, the datamodeling framework is proving itself to be a really game-changing tool

in developing data-centric view on machine learning. It is attracting attention from other leading research groups too (and also garners a real interest at OpenAI—which is the place I am spending time at)! It is very much on trajectory to become a truly influential line of work.

Again, the above results are in the end just a sliver of the impressive body of work that Andrew has produced. In fact, in addition to this machine learning focus, he *also* has a strong interest and contributions to the area of AI policy—see `https://aipolicy.substack.com/` to sample some of that (and it is really worth the read!). In the end, for Andrew, the sky is the limit—if he is passionate about something, he will find a way to engage and leave his footprint there.

Of course, this amazing productivity and breadth is not coming from nowhere. Andrew is an extremely driven person and has a truly amazing work ethics to boot. What I find very unusual about him, however, is that despite all of that *he is very down-to-earth humble, selfless and unassuming*. He is also an ultimate team player too. Always ready to step up to some ungrateful work that needs to be done. Also, always willing to share—or even cede—credit. (More than once I have seen Andrew decide to let a student that "needed the credit more" have that credit.) In fact, by now, *Andrew is a true legend in my group*, someone that my students are eager to interact with and learn from—serving in a true "secondary advisor" role for some of them. (Needless to say, I am convinced he will be a terrific "primary advisor" too!)

Let me conclude by saying that Andrew is *truly extraordinary* and *once-in-a-decade* candidate. He is a "force of nature"—fearless and unrelenting in the pursuit of the goals that truly matter to him. He embodies an unusual mix of "blue sky" optimism and incredible drive but also of sheer talent and crisp vision of what kind of impact he wants to make. (I keep thinking back to what he once told me: "Aleksander, in research I feel best when I am annoyed and puzzled by something"—that was exactly his state of mind when working on the adversarial examples and datamodels results I described above.)

All in all, I have served on numerous faculty hiring committees (including chairing some) and my advice to you would be simple: *interview and then hire him!* You simply cannot go wrong with hiring Andrew (and *will* regret if you simply pass on him). He has made lasting contributions to the field of ML already but it is clear that this is just a start. You will be extremely happy when he, inevitably, continues to do exactly that at your department.

Please do not hesitate to contact me if you need any further information.

Sincerely,

Aleksander Mądry

Aleksander Mądry
CDS Professor of Computing
MIT EECS Department