search

CATEGORIES

SHARE THIS ARTICLE

SUBSCRIBE

Get notified about new posts

Email Address

SIGN UP

Should It Be Easier to Trust Machines or Harder to Trust Humans?

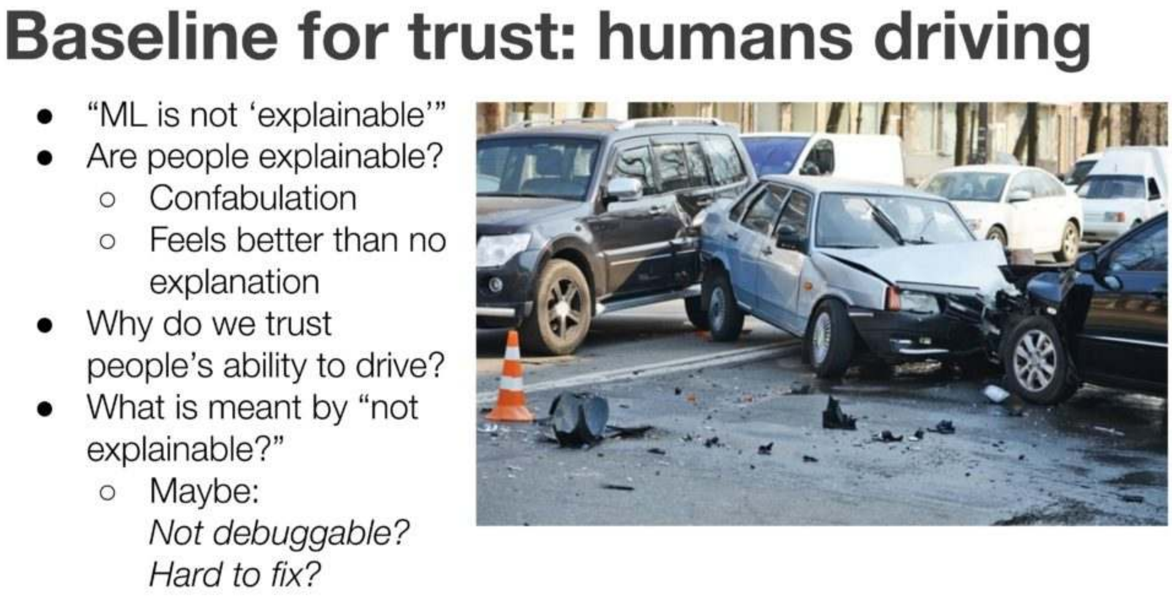
MONDAY, DECEMBER 13, 2021

MACHINE LEARNING (HTTPS://GALOIS.COM/BLOG/CATEGORY/MACHINE-LEARNING/)

Walt Woods (https://galois.com/?post_type=person&p=2639)

This blog post derived from a presentation given 2021-11-12 at a workshop for the University of Southern California's Center for Autonomy and Artificial Intelligence (https://aai.usc.edu/).

Black-box machine learning (ML) methods, often criticized as difficult to explain, can derive results with an accuracy that matches or exceeds human ability on real-world tasks. This has been demonstrated in applications such as skin lesion identification (Tschandl 2017) and telephone conversation transcription (Stolcke 2017). Still, there is a greater discomfort when it comes to using these methods in potentially dangerous scenarios that could benefit from their application, such as driving. There is a significant body of literature developing around the hypothesis that this discomfort largely stems from lacking the ability to explain or otherwise inspect the decisions coming from such ML systems.



Baseline for trust: humans driving

- "ML is not 'explainable'"
- Are people explainable?
  - Confabulation
  - Feels better than no explanation
- Why do we trust people's ability to drive?
- What is meant by "not explainable?"
  - Maybe: *Not debuggable? Hard to fix?*

(https://galois.com/wp-content/uploads/2021/12/slides-2-scaled.jpg)

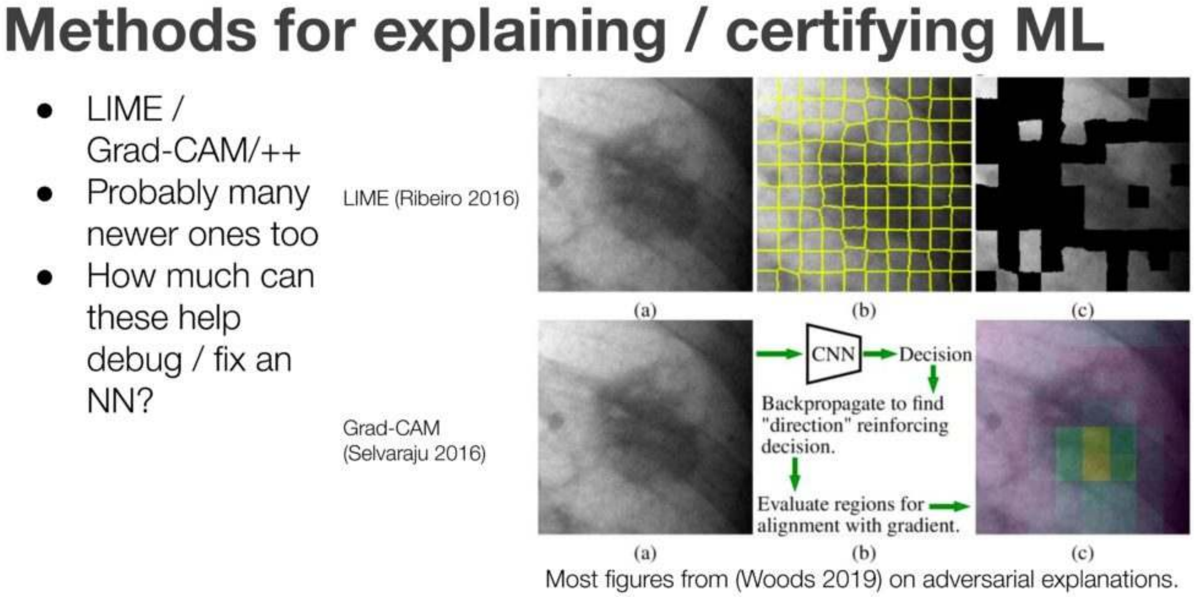
Yet, there is extant proof that we don't typically concern ourselves with such explanations: specifically, that we allow people to drive. People have been driving for over 100 years, and we more or less trust our fellow commuters to do the right thing on the road. Driver licenses are awarded without a thorough examination of the underlying explanations for all decisions made during a driving test; instead, proof of ability to do the right thing in various situations is accepted as proof that one should be allowed on the road alongside fellow drivers. Driver education classes espouse "defensive driving," since we're never quite sure what the wobbly station wagon in the lane next to us is up to, but cautionary techniques do not equate to us accurately explaining the decisions of other commuters.

Furthermore, we might claim that while we do not *have* explanations for our road team of fellow drivers, we could ask them for such an explanation in the event of an accident. This would be the claim that humans are *explainable*, and thus better suited to driving than these ML methods. A very interesting concept from psychology literature applies at this point: confabulation, when people "tell a story that is not backed up by the relevant evidence, although they genuinely regard it as a true story" (Bortolotti 2018). Bortolotti notes that people confabulate for many different reasons; one theory is that, "confabulating can have some advantages over offering no explanation because it makes a distinctive contribution to people's sense of themselves as competent and largely coherent agents" (Bortolotti 2018). This calls into question any explanation that a person might offer — we rarely have all relevant evidence, and so the accuracy of someone's explanation is notoriously difficult to assess; this can be seen from the litany of factors courts consider that affect the reliability of eyewitness testimony (Magnussen 2010).

Consequently, for the sake of discussion, let's cast aside the notion that humans are explainable. Why then do we trust them to drive or assess skin lesions? It cannot simply be a track record of success, because ML has reached the point where it can match or surpass humans on that front. Perhaps it has more to do with our trust that humans are teachable through verbal and non-verbal communication. For example, a driving instructor could turn to a novice driver who's being a bit careless and say, "I'd like you to consider this as a dangerous driving situation because there's a pedestrian in the crosswalk," and trust that the instructor and driver's understanding of the base concepts of "dangerous", "pedestrian," and "crosswalk" were sufficiently aligned, and trust that the importance of the message were sufficiently conveyed, and trust that the driver would recognize a similar situation in the future, and trust that the novice driver would react differently the next time to account for the danger. All of this has very little to do with explaining or debugging the driver's behavior and much more to do with an ability to enact desired change through the exploitation of shared concepts (pedestrian, crosswalk) and values (avoiding danger). In other words, communicating that *danger = crosswalk.contains(pedestrian)*.

That seems reasonable, right? The problem is, it's for a very simple situation with bold, easily identified variables. That is, we have language which robustly conveys all pertinent aspects, while discarding irrelevant information. For example, the word "pedestrian" implies any person walking or standing around, regardless of what color shirt they're wearing, or if they're wearing a hat or a scarf, etc. We also implicitly believe that the novice driver has the ability to identify such an abstract concept in a variety of sizes and angles, and under different lighting, weather, or other visibility-impairing obstruction conditions. The real world is complicated, and what allows us to convey instructive logic effectively is the ability to harness these robust concepts when specifying a simple, logical relation between them.

The disconnect in ease of communication between teaching a new concept to a person versus an ML algorithm isn't in communicating the formal logic of the concept, *danger = crosswalk.contains(pedestrian)*. The challenge is learning robust definitions for the variables outside of the formal framework. The problem then becomes debugging and showing the robustness of ML decisions, rather than arbitrarily "explaining" them.



Methods for explaining / certifying ML

- LIME / Grad-CAM/++
- Probably many newer ones too
- How much can these help debug / fix an NN?

LIME (Ribeiro 2016)

Grad-CAM (Selvaraju 2016)

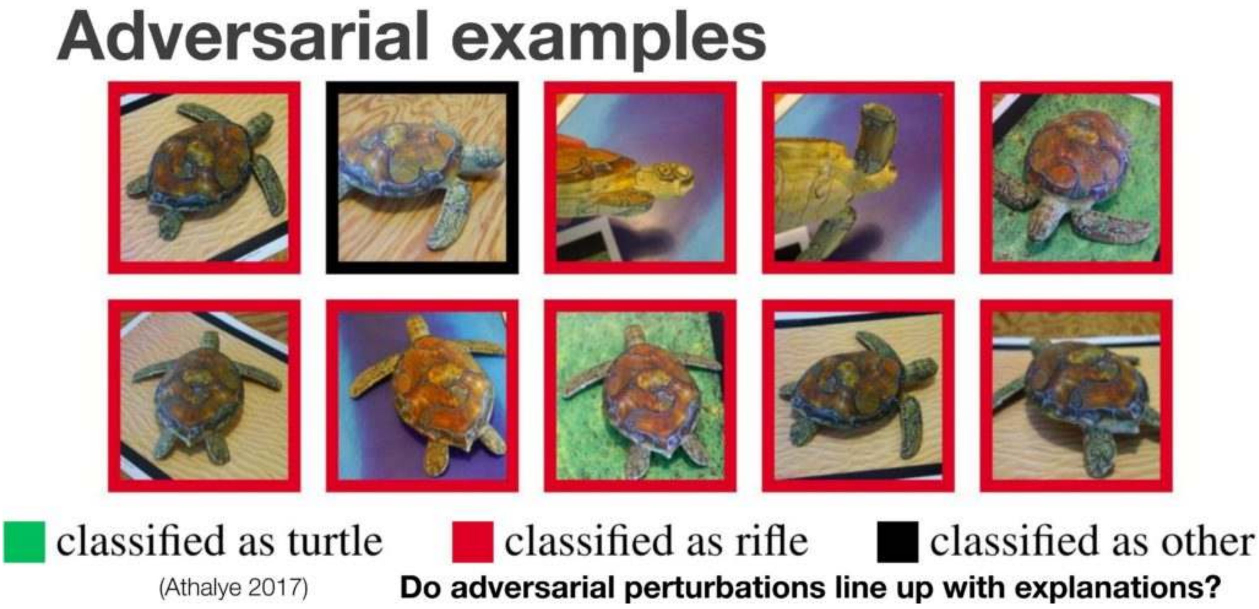
(a)
(b)
(c)
CNN
Decision
Backpropagate to find "direction" reinforcing decision.
Evaluate regions for alignment with gradient.
(a)
(b)
(c)

Most figures from (Woods 2019) on adversarial explanations.

(https://galois.com/wp-content/uploads/2021/12/slides-3-scaled.jpg)

With that established, we can talk about existing methods of explanation. LIME (Ribeiro 2016) and Grad-CAM (Selvaraju 2016) are two well-known means of gathering insight from neural network decisions. Roughly, these techniques mask out an image such that what remains are the parts of the image that were most relevant to the decision-making process. As seen above, in the case of identifying lung nodules from x-rays, LIME indicates that various parts of the image were used, but particularly the middle-right region of the image. It's hard to know what to do with this information to improve the network. Similarly, for Grad-CAM, the middle-right region of the image was most important. In the Grad-CAM case, though, it appears to be highlighting the lower edge of the nodule, which is encouraging.

Neither of these mechanisms provide actionable insight into how one might better teach these networks, and neither of these mechanisms help us decide the robustness of these decisions.



Adversarial examples

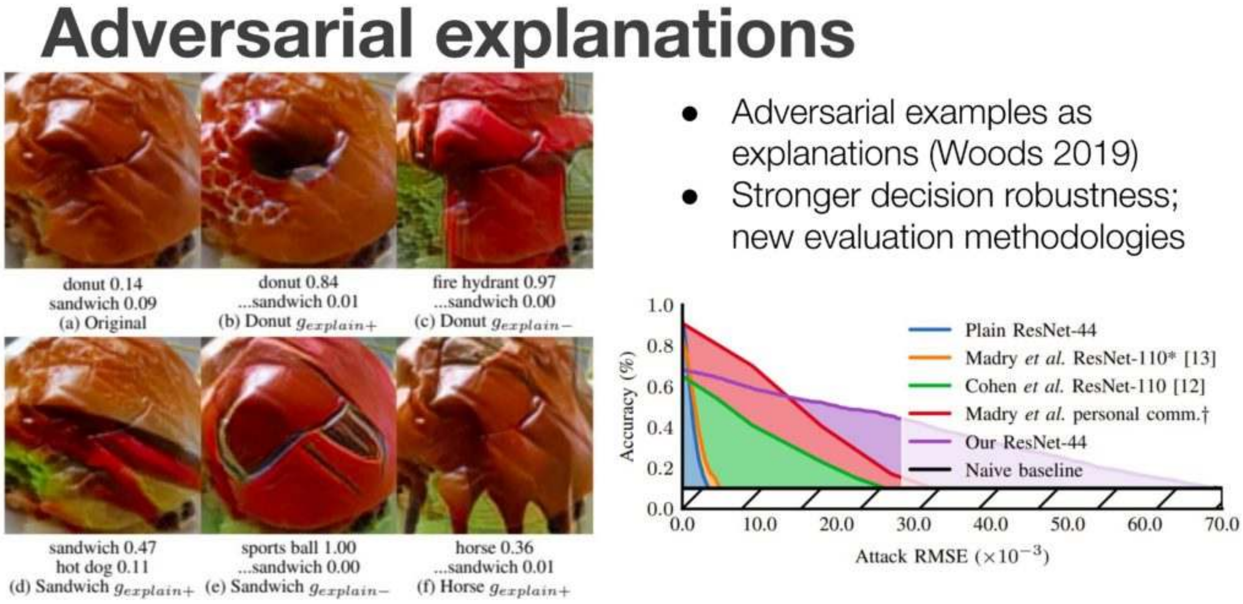
🟩 classified as turtle 🟥 classified as rifle ⬛ classified as other

(Athalye 2017)

Do adversarial perturbations line up with explanations?

(https://galois.com/wp-content/uploads/2021/12/slides-4-scaled.jpg)

Adversarial examples are a fascinating attack against machine learning algorithms, and neural networks in particular. In short, neural networks have a large amount of gain on their inputs; most images can be manipulated such that a change to an image, which is imperceptible to a human observer, completely changes the output of the neural network. What's shown above is a wonderful example from (Athalye 2018): they fabricated a 3-dimensional turtle and painted it such that Google's state-of-the-art image classification network almost always considered the object to be a rifle instead of a turtle, regardless of presentation angle or lighting conditions. This demonstrates that adversarial examples are not only a significant real-world threat, but also that most extant networks make decisions in a non-robust manner. Since such a small change to the input completely changes the decision, we could say that the LIME and Grad-CAM algorithms have an extremely limited domain of validity. They are valid for a particular input, but not the neighborhood around it. In other words, they won't help us move toward networks that can be taught new logical concepts that rely on a robust foundation.



Adversarial explanations

- Adversarial examples as explanations (Woods 2019)
- Stronger decision robustness; new evaluation methodologies

donut 0.14
sandwich 0.09
(a) Original

donut 0.84
...sandwich 0.01
(b) Donut $g_{explain+}$

fire hydrant 0.97
...sandwich 0.00
(c) Donut $g_{explain-}$

sandwich 0.47
hot dog 0.11
(d) Sandwich $g_{explain+}$

sports ball 1.00
...sandwich 0.00
(e) Sandwich $g_{explain-}$

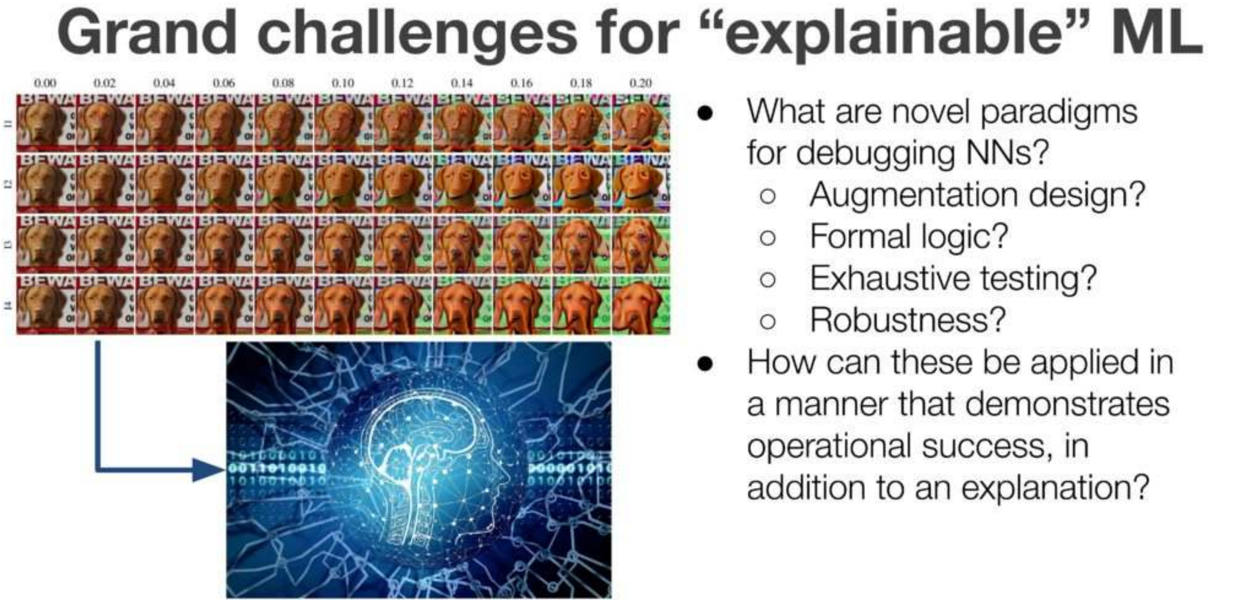
horse 0.36
...sandwich 0.01
(f) Horse $g_{explain+}$

Accuracy (%)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
Attack RMSE ($\times 10^{-3}$)
Plain ResNet-44
Madry et al. ResNet-110* [13]
Cohen et al. ResNet-110 [12]
Madry et al. personal comm.†
Our ResNet-44
Naive baseline

(https://galois.com/wp-content/uploads/2021/12/slides-5-scaled.jpg)

In contrast, recent work has proven successful at leveraging adversarial attacks themselves as part of the explanation process (Woods 2019). By conditioning the network with a fairly straightforward Lipschitz regularization, adversarial attacks accentuate key features of the underlying concept, rather than appearing as random noise. An interesting corollary of this method of training networks is that these networks are provably more robust than standard machine learning networks — they demonstrate lower accuracy on clean data, but as an attacker is allowed to modify the input data, their accuracy quickly overtakes that of networks without such robustness. On problems with a large domain, like ImageNet, these networks scored an Accuracy-Robustness area of 2.4x higher than the next-most-robust networks.

While it would be nice if this were the be-all-end-all solution for robustness, the lower accuracy on clean data is significant. Furthermore, while techniques like this help to ensure the robustness angle for underlying concepts needed for more teachable ML, they don't necessarily help us debug and correct the network. We can see biases in the data, such as the usual angle at which fire hydrants are seen, but beyond a small amount of active learning, or changing the training data to be more diverse, there is no direct way to teach the network. Adversarial explanations are a great step toward debuggable, robust ML, but they require additional logical abstraction. On their own, they're not enough to leverage the instructional communication that we expect and are comfortable with when interacting with other people.



Grand challenges for "explainable" ML

0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20

- What are novel paradigms for debugging NNs?
  - ○ Augmentation design?
  - ○ Formal logic?
  - ○ Exhaustive testing?
  - ○ Robustness?
- How can these be applied in a manner that demonstrates operational success, in addition to an explanation?

(https://galois.com/wp-content/uploads/2021/12/slides-6-scaled.jpg)

We trust humans more than we trust ML on many tasks. Despite a growing track record of ML solutions that demonstrate excellent performance, there's a missing element that has many people stopping short of being willing to accept ML-based decisions. It's unlikely that this missing element is an explanation, as we trust humans to do innumerable dangerous tasks, including driving, and humans lack the quality of being truly explainable. Instead, I've proposed that we consider making networks teachable, using similar communication primitives that we use with people: simple logic that relates robust concepts. Computers are already great at simple logic — it's what has made them so useful in the modern era. On the other hand, teaching them to identify concepts robustly is quite difficult. Techniques like adversarial explanations help, but they are only a piece of the puzzle.

Looking forward, we expect the development of novel paradigms for teaching and debugging neural networks. These might take the shape of clever augmentation design coupled with few shot learning, or perhaps some form of differentiable formal logic. It might take the shape of a clever exhaustive testing mechanism that can nudge networks toward robust logics, or perhaps something like adversarial explanations that encourage robustness, but in a more clever way. Regardless of the mechanism, ML will need to advance toward more teachable and robust networks to show that they can deal with the unpredictable complexities of real-world use while retaining their superhuman performance.

For further discussion, feel free to reach out to Walt at waltw@galois.com (mailto:waltw@galois.com) *directly.*

References

[Athalye 2018] Athalye, Anish, et al. "Synthesizing robust adversarial examples." *International conference on machine learning*. PMLR, 2018.

[Bortolotti 2018] Bortolotti, L. Stranger than Fiction: Costs and Benefits of Everyday Confabulation. Rev.Phil.Psych. 9, 227–249 (2018). https://doi.org/10.1007/s13164-017-0367-y (https://doi.org/10.1007/s13164-017-0367-y)

[Magnussen 2010] Magnussen, Svein, et al. "Beliefs about factors affecting the reliability of eyewitness testimony: A comparison of judges, jurors and the general public." *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 24.1 (2010): 122-133.

[Ribeiro 2016] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." *arXiv preprint arXiv:1606.05386* (2016).

[Selvaraju 2016] Selvaraju, Ramprasaath R., et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization.” *Proceedings of the IEEE international conference on computer vision*. 2017.

[Stolcke 2017] Stolcke, Andreas, and Jasha Droppo. “Comparing human and machine errors in conversational speech transcription.” *arXiv preprint arXiv:1708.08615* (2017).

[Tschandl 2019] Tschandl, Philipp, et al. “Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study.” *The Lancet Oncology* 20.7 (2019): 938-947.

[Woods 2019] Woods, Walt, Jack Chen, and Christof Teuscher. “Adversarial explanations for understanding image classification decisions and improved neural network robustness.” *Nature Machine Intelligence* 1.11 (2019): 508-516.

PREVIOUS (HTTPS://GALOIS.COM/BLOG/2021/12/THE-2021-GALOIS-BALLOWEEN-WORKSHOP/)

NEXT

Most Recent Tech Talk

Title John Launchbury: The Trajectory of AI (https://galois.com/blog/2023/12/the-trajectory-of-ai/)

Date Friday, December 01, 2023 **Time** 11:00 am

Speaker John Launchbury

Location Portland, OR

About "In 2015 I started talking about Three Waves of AI as a framework for understanding the new burst of machine learning developments that were taking place, and to put DARPA I2O’s research portfolio into context. Eight years later, ChatGPT

(HTTPS://GALOIS.COM/BLOG/2023/12/THE-TRAJECTORY-OF-AI/)

Galois News

Galois Releases the Swanky Suite of Rust Libraries for Secure Computation (https://galois.com/news/galois-releases-the-swanky-suite-of-rust-libraries-for-secure-computation/)
PRESS RELEASE

Galois Releases CAMET Base Pack 1.6.1 with Enhanced Capabilities and Stability Improvements (https://galois.com/news/galois-releases-camet-base-pack-1-6-1-with-enhanced-capabilities-and-stability-improvements/)
PRESS RELEASE

(HTTPS://GALOIS.COM/NEWS/)

Portland, OR

421 SW 6th Avenue, Suite 300
Portland, Oregon 97204
(https://www.google.com/maps/place/Galois,+Inc./@45.520811,-122.678081,17z/data=!4m6!1m3!3m2!1s0x54950a04159ece0f:0x36857895c75e27d7!2sGalois,+Inc.!3m1!1s0x54950a041

Arlington, VA

901 N Stuart Street, Suite 501
Arlington, Virginia 22203 (https://goo.gl/maps/pxFK95q48t32)

Minneapolis, MN

111 Third Avenue South, Suite 350
Minneapolis, MN 55401 (https://maps.app.goo.gl/csQrxYhmMPHDXLZJA)

Dayton, OH

444 E 2nd Street
Dayton, Ohio 45402 (https://goo.gl/maps/cRzPpKEF6eD2)

T 503.626.6616 (tel:15036266616)
F 503.350.0833

contact@galois.com (mailto:contact@galois.com)