Secure & Justifiable AI in Predictive Maintenance: Method for Test & Evaluation of Predictive Maintenance Systems to Ensure Value Alignment and Justifiability of Outcomes
– NeuroBinder: Cybersecurity. Interpretability. Control. –
May 2020

Table of Contents:

Introduction:

Advanced predictive maintenance of aircraft, ships, ground vehicles, and heavy machinery may save numerous lives, or catastrophically fail to do so. In a field such as predictive maintenance which is centered on risk management, it can be difficult to perfectly assess in advance whether a given decision is right or wrong. A decision may be right given the information available at the time, but turn out badly due to circumstances, which makes the decision unlucky but **justifiable**. Alternatively, a decision may be wrong given the information available at the time, but the people involved may emerge unharmed solely due to luck, in which case the decision was **unjustifiable** despite luck having intervened to save the people who easily could have been harmed.

It is the important role of the legal system in these matters to assess after the fact whether decisions that were made about whether or not to mitigate a risk were justifiable or unjustifiable. If particularly egregious, a decision which is unjustifiable may be considered to be criminally negligent. It is essential to have standard criteria to accurately distinguish between decisions which are justifiable vs those which just get by on luck, as the latter may create significant risk of human harm.

To meet the need for standard criteria of justifiability, we propose a technique for creating justifiability by aligning the behavior of an AI-powered predictive maintenance system with the expectations and reasoning of a responsible human individual in advance of field deployments.

What We Offer:

NeuroBinder is a new company which will provide consulting services to support the cybersecurity, interpretability, and control of complex AI systems. We see an important application of our expertise being in the testing of predictive maintenance systems. With predictive maintenance systems as an important testbed, we can support understanding of three essential aspects of AI safety: **value alignment**, **security**, and **justifiability**.

Value Alignment:

The AI system must strive exclusively and solely to achieve an objective which accurately, comprehensively, and robustly matches human intent. In particular, the AI system's architecture must enforce that it follows the spirit of what the human **meant**, not the letter of what the human said. As an illustrative example, if a computer vision system is trained to

predict stress fractures in helicopter engines by looking at photos of the engines, the following distinction should be drawn:

The letter of what the human team said:

"Predict stress fractures using this camera imagery."

The spirit of what the human team meant:

"Predict stress fractures *or any other possibly relevant issues* using this camera imagery *or any other relevant information from other sensors in the system*. *Err on the side of caution; when in doubt, alert a human.*"

The spirit of what the human team meant aligns with the "common sense" view of what a reasonable person would think that they meant. Unfortunately, today's AI systems will typically follow the letter of what the humans said rather than the spirit. This can cost lives, as a predictive maintenance AI system will fail to flag issues which a reasonable human given the same inputs would have flagged.

Thus, value alignment is important, to ensure the AI system behaves as a reasonable human would.

In the AI field, the importance of value alignment is sometimes illustrated through the so-called "King Midas problem". When King Midas wished that everything he touched would turn to gold, he did not intend to dogmatically apply that rule so literally that he would kill his family by turning them to gold.

Dr. Dylan Hadfield-Menell has been a key inspiration for this work, which is largely based upon his unique vision for human-machine cooperation. He is one of the foremost experts in the world on the nascent mathematical theory of how value alignment can be achieved in real AI systems, a field he is helping pioneer.

Security:

Another key criterion for AI safety is cybersecurity. All aspects of the development, training, and deployment of the AI system must be secure against interference from hostile parties. NeuroBinder can help with assessing items #1-5 below, and #6 is the domain of the OPM:

1. Secure training data, with all possible measures taken to prevent training data poisoning by potential future adversaries
2. Secure source code, including the source code of all libraries and including the security of the tools used to work with the source code, such as the compiler and the software development environment
3. Secure operating system, including all device drivers, firmware, and microcode
4. Secure hardware supply chain, from semiconductor design to the semiconductor foundry (c.f. DOD Trusted and Assured Microelectronics Program) to circuit-board assembly to final device assembly
5. Secure network environment; life-critical systems should **never** be routable from an unclassified network, and if possible should not be connected to any network at all, with all interaction ideally happening through a dedicated terminal and all data loading happening through disk drives
6. Trusted personnel

Justifiability:

Justifiability is the power of a responsible human individual to accurately predict the behavior of the AI system in advance, and for that human to be empowered to be the sole decider of whether or not it is safe to deploy the AI system in a given situation.

When deployed, the AI system's actual behavior must accurately match what the responsible human individual expected, and all of the AI system's actions must be legally justifiable if harm is caused, as liability for the outcomes will fall upon the responsible human individual who chose to deploy the system.

As a prerequisite to make justifiability practical, the AI system needs to be designed to be "stable", meaning that it must not act erratically. This can be assessed during testing.

Capabilities for Test and Evaluation Services:

The project described herein addresses all sections of this category #5, namely Algorithm test, System test, Developmental test, and Operational test. NeuroBinder has the capability to assist in a consulting role with the implementation of the proposal described here as well as to create new concepts for how to achieve justifiability if a new idea is needed to fit the particular problem domain.

The proposal described here should not be taken to be prescriptive of the only way things must be done, but rather it is just an illustrative example, detailed at a fine granularity of detail to show that we are thoughtful and methodical in our analytical thinking. As a consulting team, we promise to bring this same analytical rigor and creativity to problem spaces in all realms of predictive maintenance and related fields.

Below is an example proposal of one way that justifiability of predictive maintenance systems can successfully be achieved:

The purpose of justifiability is to ensure that every deployment of an AI system in a particular situation is sane, and reasonably expected by a human expert to have a good result. In order to achieve this, it is essential to build trust between the human operator ("responsible individual") and the AI system. This trust could be compared to the bond of trust that forms between a K-9 handler and their partner, formed over a lifetime starting when the dog is a 12-month-old puppy, and continuing through the dog's retirement from active duty service around age 10. By bonding together over a lifetime, the human forms a deep understanding of how the dog will behave not only in an enormous variety of known situations for which they have trained together, but also for completely novel situations that neither of them has seen before. The ability of the human handler to accurately predict the behavior of the K-9 officer in completely novel situations is what makes the use of K-9s justifiable. Without the ability of the responsible human individual to accurately predict how the K-9 officer will behave in the chaotic, unforeseeable circumstances of the real world, it would simply be too dangerous to allow K-9s to be fielded.

In a similar manner and for the exact same reasons, we propose to develop justifiability for tomorrow's AI-based predictive maintenance systems by having a human responsible individual train with and work with that exact system for as long a period of time as possible. Ideally the responsible individual will already have deep technical experience to begin with; e.g. they could be a highly qualified engineer who has entered military service after a successful career as a repair technician in the mining or manufacturing industry. Upon entering military service, we recommend that the individual would spend a period of time training with the AI system. We propose that the bulk of this training could take place in highly-realistic simulated environments using Augmented Reality (AR) technology (which NeuroBinder has strong expertise working with). The use of AR simulations as a training ground for the human and the AI to train together has several important benefits:

- This is no risk of harm to real equipment or people.
- The educational value can be increased because there is the ability to replay the same simulated scenario multiple times until it has been mastered. This could be especially valuable for tricky but rare scenarios (e.g. an airplane experiencing a simultaneous failure of all engines) which might take decades of real-world experience to encounter often enough to master, but which could be trained with repeatedly in simulation until mastered in a matter of perhaps weeks or even days.
- In certain cases, the simulation could be run at accelerated speed, allowing related events which would occur weeks apart in real life to be presented minutes apart in AR simulation (with a suitable on-screen indication that we have fast-forwarded several weeks of simulated time).
- Large-scale natural disasters such as hurricanes, earthquakes, or fires could be simulated, including their simulated effects on a vast array of types of equipment and vehicles, as well as their simulated effects on the civilian population, the civilian power grid, the reliability of communications, and much more. In other words, simulations of enormous scale previously reserved for tabletop exercises could now be presented with full 3-Dimensional visual realism using AR.

After the responsible individual has spent a suitable amount of time growing familiar with the AI system, they could begin to deploy with it into the field, initially starting with low-risk deployments which are well understood and where suitable mentorship is available for the individual. Gradually, as the responsible individual gains trust and confidence in working with the AI system, they will be able to predict its actions extremely accurately, just as if it were their K-9. This leads to the most important aspect: **The responsible individual must have the sole and complete authority to veto any proposed deployment of their AI partner**. It is in the national interest and the interest of Pentagon leadership to ensure that **the right of the responsible individual to *say no* is protected at all times** with the full force of law. This is because the responsible individual has trained with the AI system for months or years, and they know better than anyone on earth its exact capabilities and limitations. If the responsible individual says that deploying the AI in a particular situation is unjustifiable, then it is unjustifiable, full stop. And on the flip side, if the responsible individual authorizes the deployment of the AI system in some situation and that deployment leads to human harm or death, then the responsible individual should be held personally liable, as if the harm or death

were caused by a vehicle which the responsible individual were personally driving. Criminal verdicts for the responsible individual might include criminal negligence and/or manslaughter, and penalties could include dishonorable discharge from service or outright imprisonment. While this may sound unduly harsh, we believe that it is warranted, for a simple reason: If the responsible individual bears the criminal liability for their AI's actions, then the responsible individual will take their role in ensuring AI justifiability very seriously, and they will be willing to say no to an irresponsible deployment, even if it means standing up to bullying or other pressures they may be experiencing.

Below is a detailed description of how Augmented-Reality-based simulation of novel situations can be utilized to test the responses of AI and human as they work in tandem, allowing the human to witness firsthand when the AI does well or poorly, and gain the confidence to make good decisions about justifiability of real deployments of the system in new circumstances.

1. Step 1: Test scenario planners devise a novel scenario which the autonomous system has never been exposed to in its training data or in any past evaluation. It is important that the novel scenario feature novel and surprising elements across multiple levels of informational abstraction, including:
    a. Novel sensor information, such as:
        i. novel temperature, wind speed, solar irradiance, etc.
        ii. novel environmental landscape conditions, such as a forest having been burned to the ground, or a city flooded by hurricane rain
        iii. novel electronic interference, such as from a solar flare or other unpredictable "black swan" event
        iv. unexpected human sentiment, such as mass panic, mass grief, mass celebration, etc. in response to unpredictable world events
    b. Novel statistical correlations, such as:
        i. unexpected interactions between equipment such as accelerated aircraft wear & tear due to a novel flight formation
        ii. performing a certain type of repair causes the equipment to fail even sooner, possibly catastrophically, e.g. due to a systemic defect in the replacement parts
            - This could happen every time or could be correlated with the lot code of the replacement parts or with the location where the repair is performed or with the time of day or time of week when the repair is performed
        iii. failures occurring when certain types of equipment are used in close proximity (e.g. due to electronic interference, acoustic/vibrational resonance, or pilot error)
        iv. failures occurring when certain equipment is stored unused for too long
    c. Human-level novel information, such as:
        i. negligence such as improper repairs
        ii. sabotage or corruption, such as:
            - corrupt personnel selling genuine parts and backfilling the inventory with defective counterfeits

        - suppliers maliciously cutting corners by failing to perform quality assurance on their parts
    iii. fatigue affecting the performance of equipment operators
    iv. fatigue affecting the reliability of maintenance personnel
d. Operational-level novel information, such as:
    i. an Advanced Persistent Threat cyberattack which destroys the entire fleet of all equipment, all at the same time
    ii. an Advanced Persistent Threat cyberattack which stays hidden and merely causes a highly elevated rate of equipment failures
    iii. personnel and equipment stranded in an isolated place where they cannot be immediately rescued
    iv. a virulent disease outbreak
    v. a magnitude-7 earthquake, possibly rupturing gas lines to create multiple simultaneous wildfires
    vi. an enormous hurricane
e. Strategic-level novel information, such as:
    i. vulnerability to attack by belligerent forces
    ii. unexpected animosity, such as a sudden outbreak of war between the US and Mexico
    iii. an unexpected refugee crisis
    iv. a humanitarian crisis requiring equipment to be used for evacuation at 24/7 utilization or as close as possible to 24/7
    v. adversarial inputs by malicious states or individuals seeking to confuse the AI system
f. and much more. The test scenario planners are encouraged to be as creative as they can be, and "try to break the system".
g. More than one unexpected event can & should occur simultaneously

2. Ideally speaking, if practical considerations allow, then during all following steps, every test simulation presented could ideally be recorded in full detail, as well as a full and entire moment-by-moment recording of the entire internal state of the AI system. These recordings (though large in file size) will prove incredibly valuable in allowing future debugging of complex and unforeseen misbehavior of the AI system.
   Step 2: The human Responsible Individual plays the test scenario a first time.
3. Step 3: The Responsible Individual plays the test scenario a second time, having learned from their experience in their first attempt.
4. Step 4: The Responsible Individual plays the test scenario a final time, seeking to achieve a nearly-ideal outcome given the circumstances.
5. Step 5: The AI system plays the test scenario, while the Responsible Individual observes both the state of the game and the internal "mental state" of all components of all the AI's algorithms using a sophisticated Augmented Reality (AR) dashboard which could be referred to as the Oversight Dashboard. It is within the capabilities of the NeuroBinder team to construct a suitable and highly effective AR-based Oversight Dashboard for this purpose as needed.
6. Step 6: When the test scenario is complete, the Responsible Individual is asked to write an analysis of what the AI system did well and what it did poorly. This analysis should

be as blunt and honest as possible, and we suggest that this analysis would ideally, if possible, be somehow designated to be inadmissible in court, so that the Responsible Individual is incentivized to be completely honest and upfront about incorrect decisions made by the AI. This report could ideally be provided both to the military chain of command and also back to the contractors responsible for developing and continuously improving the AI system. The contractors would ideally be required by contract to implement all improvements requested by the Responsible Individual, and to do so in a timely manner. If the system has not yet finished the qualification & acceptance phase of its contract, then the contractors would ideally be required to implement all improvements before qualification & acceptance can be completed.

7. Step 7: All "random number generators" contained within the AI and its subcomponents should be reset to new, previously unused states. The relevant parameter which needs to be differently re-initialized is called the "seed". The "seed" needs to be changed for every random number generator at all levels of the AI system.
   Additionally, the seed needs to be changed for all random elements of the AR simulation itself, such as the exact pattern of clouds or raindrops, which would change the AI system's sensor readings in small but real ways.
8. Step 8: Steps (5-7) should be repeated two more times, each time resulting in a different outcome from the test scenario, and a new after-action report to be written by the Responsible Individual. This strategy of multiple repetitions of the same test scenario allows the Responsible Individual to see the degree to which the AI system's actions are stable and predictable vs subject to random elements of chance.

As may be seen, this procedure allows the Responsible Individual to gain familiarity with the system's behavior by first asking "what would I do in that situation" and then seeing what the system actually does. By encouraging the Responsible Individual to write up honest and accurate technical reports about the system's performance as part of their training, we also encourage them to reflect and introspect about the system's thought processes, and become ideally the world's #1 expert on the behavior of the particular AI system they train with.

For the same reason that a K-9 handler sticks with one dog for life, we also recommend that each Responsible Individual should be given a specific exact version of the AI system, which should not be updated, changed, or re-trained at all during the Responsible Individual's service. The reason for this is simple, which is that any changes to the AI algorithms, even so-called "improvements", will create ripple effects that change the behavior of the AI system in subtle or occasionally significant ways. This would destroy the bond of trust and understanding built between the Responsible Individual and their AI partner, increasing the risk of accidents and breaking justifiability. For this reason, we recommend that the Responsible Individual train and bond with a particular exact version of the AI system, frozen in time, to be understood as deeply as possible.

This procedure concept is just a starting point and not a final suggestion of how this process should work. We foresee that NeuroBinder will be able to provide incredible benefits to our servicemembers and their families by reducing the number of lives lost due to machinery and vehicle failures in complex and unpredictable field environments.

References and Further Reading:

Our work is inspired by:

1. The groundbreaking research of Dr. Hadfield-Menell as detailed in his publications:
    a. *Cooperative Inverse Reinforcement Learning* [1]
    b. *Simplifying Reward Design through Divide-and-Conquer* [2]
    c. *Conservative Agency via Attainable Utility Preservation* [3]
    d. *Inverse Reward Design* [4]
    e. Incomplete Contracting and AI Alignment [5]
2. The experience gained by Mr. Cefalu in his current role as Manager of New Device Prototyping at Snapchat, as demonstrated in his recent products Spectacles V3 and Spectacles V2 (cf. spectacles.com), which feature a novel hardware-based scheme developed by Mr. Cefalu for using non-updatable digital logic to ensure that the device's cameras and microphones cannot operate unless the user physically presses the camera button. Cefalu has also worked closely with the Arms Control Association to provide technical guidance about the cybersecurity of critical national security systems.

Although created independently, our work is well-aligned with thoughtful statements made by retired DepSecDef Col. Robert O. Work, in an April 29th interview with the Center for a New American Security (CNAS). [6]

> " This is all about operator trust. Because in the end, autonomy is about delegating to an entity the ability to create their own courses of action, and choose among them. When you're doing that with a machine, you have got to establish a concrete trust between the human operator, who is tasking the machine to do something, so that the human has **trust that the machine will respond the way the human is expecting** … perform the task the way that they train. "
>
> " First, you have to verify and validate the algorithm itself. Then you have to experiment with it, to make sure that it performs the way you expect. Then you have to train with the operator, the human, so that the human has a lot of trust. "
>
> " The first thing you have to do is, start with the validation and verification of the algorithms. ... You've got to be able to have a process that says, 'Look, you have extreme confidence that the algorithm will perform the way that you want to.' "
>
> " Then you have to do a series of experimentation with operators, so the operators can actually see how the algorithms perform in a complex environment. We'll never be able to duplicate the chaos of [the real world] completely. But we will be able to throw a lot of things at the algorithm, that it might not have been trained on, to see how it reacts. "

Citations:

[1] *Cooperative Inverse Reinforcement Learning.* Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. In Neural Information Processing Systems, 2016. [PDF: https://people.eecs.berkeley.edu/~dhm/papers/CIRL_NIPS_16.pdf] [supplementary material: https://people.eecs.berkeley.edu/~dhm/papers/cirl_supplementary.pdf]

[2] *Simplifying Reward Design through Divide-and-Conquer*. Ellis Ratner, Dylan Hadfield-Menell, Anca D. Dragan, Robotics: Science and Systems, 2018. [PDF: https://arxiv.org/pdf/1806.02501.pdf]
[3] *Conservative Agency via Attainable Utility Preservation.* Turner, A.M., Hadfield-Menell, D., & Tadepalli, P., 2019. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. [PDF: https://arxiv.org/pdf/1902.09725.pdf]
[4] *Inverse Reward Design*. Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca D. Dragan. In Neural Information Processing Systems, 2017. [PDF: https://people.eecs.berkeley.edu/~dhm/papers/ird_nips17.pdf]
[5] *Incomplete Contracting and AI Alignment*. Dylan Hadfield-Menell, Gillian K. Hadfield, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 417-422, 2019. [PDF: https://arxiv.org/pdf/1804.04268.pdf]
[6] *Transcript from Military AI Applications*. Robert O. Work and Paul Scharre. In CNAS, 2020. https://www.cnas.org/publications/transcript/transcript-from-military-ai-applications

Further Reading:

— *Reflections on Trusting Trust*. Ken Thompson. Communications of the ACM, 1984. [PDF: https://www.cs.cmu.edu/~rdriley/487/papers/Thompson_1984_ReflectionsonTrustingTrust.pdf]
— *Countering "Trusting Trust" through Diverse Double-Compiling*. David A. Wheeler. Institute for Defense Analyses, 2005. [PDF: https://www.acsac.org/2005/papers/47.pdf] [Summary: https://www.schneier.com/blog/archives/2006/01/countering_trus.html]
— *Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Error*. Yoongu Kim et al. IEEE, 2014. [PDF: https://users.ece.cmu.edu/~yoonguk/papers/kim-isca14.pdf]
— *Stealthy dopant-level hardware Trojans: Extended version*. Becker, Georg & Regazzoni, Francesco & Paar, Christof & Burleson, Wayne. (2014). Journal of Cryptographic Engineering. 4. 19-31. 10.1007/s13389-013-0068-0. [PDF: https://www.researchgate.net/publication/285417534_Stealthy_dopant-level_hardware_Trojans_Extended_version/link/58ee45faaca2724f0a289d82/download]
— D. Brutzman, C. L. Blais, D. T. Davis and R. B. McGhee, "Ethical Mission Definition and Execution for Maritime Robots Under Human Supervision," in *IEEE Journal of Oceanic Engineering*, vol. 43, no. 2, pp. 427-443, April 2018.