The wall of safety for AI: approaches in the confiance.ai program

Bertrand Braunschweig[1], Rodolphe Gelin[2], François Terrier[3]

[1]Institut de Recherche Technologique SystemX
2, boulevard Thomas Gobert – Bâtiment 863 F-91120, Palaiseau, France
[2]Renault Group TCR, 1 avenue du Golf, 78084 Guyancourt, France,
[3]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[1]bertrand.braunschweig@ext.irt-systemx.fr, [2]rodolphe.gelin@renault.com, [3]francois.terrier@cea.fr

Abstract

AI faces some « walls » towards which it is advancing at high pace. Apart from social and ethics consideration, there are walls on several subjects very dependent but gathering each some considerations from AI community, both for use, design and research: trust, safety, security, energy, human-machine cooperation, and « inhumanity ». Safety questions are an particularly important subjects for all of them. The Confiance.ai industrial program aims at solving some of these issues by developing seven interrelated projects that address these aspects from different viewpoints and integrate them in an engineering environment for AI-based systems. We will present the concrete approach taken by confiance.ai and the validation strategy based on real-world industrial use cases provided by the members.

The walls of AI and their relation with safety

Artificial intelligence is advancing at a very fast pace, both in terms of research and applications, and is raising societal questions that are far from being answered. But as it moves forward rapidly, it runs into what we call the five walls of AI, walls that it is likely to crash into if we don't take precautions. Any one of these five walls is capable of halting its progress, which is why it is essential to know what they are and to seek answers in order to avoid the so-called third winter of AI, a winter that would follow the first two in the years 197x and 199x, during which AI research and development came to a virtual standstill for lack of budget and community interest. The five walls are those of trust, energy, safety, human interaction and inhumanity. They each contain a number of ramifications, and obviously interact.

There are different opinions on this matter. The paper (Bengio et al. 2021) by Yoshua Bengio, Yann LeCun and Geoffrey Hinton, written after their collective Turing Award, provides insights into the future of AI through deep learning and neural networks without addressing the same topics; the 2021 progress report of Stanford's 100-year longitudinal study (Littman et al. 2021) examines AI advances to date and presents challenges for the future, very complementary to those we discuss here; the recent book by César Hidalgo (2021) looks at how humans perceive AI (and machines); the book "Human Compatible" by Stuart Russell (2019), is interested in the compatibility between machines and humans, a subject we treat differently when we talk about the interaction wall.

Trust and safety

If people do not trust the AI systems they interact with, they will reject them. Several organizations are trying to provide definitions of what is trust in artificial intelligence systems, it has been the main subject of the group of experts mobilized by the European Commission (whose work is all done in the "trustworthy AI" perspective) (EC 2019). The international standardization organization, ISO (2020a, 2020b), considers about eleven different objectives, with ramifications related to Trustworthy AI: fairness, security, safety, privacy, reliability, transparency/explainability, accountability, availability, maintainability, integrity, duty of care, social responsibility, environmental impact, availability and quality of training data, AI expertise. *This is probably not a definitive list of the dimensions of the "Trust" and all these terms would require a precise definition and the development of a dedicated ontology to identify the meaning and*

relations among them, in particular for its relations with "safety". This point motivate some activities in Confiance.ai to build a taxonomy gathering inputs from all identified sources. However, as this is still not stabilized we use them here with their inherent ambiguities.

Trust, especially in digital artifacts of which AI is a part, is a combination of technological and sociological factors. Technological, such as the ability to verify the correctness of a conclusion, the robustness to perturbations, the handling of uncertainty etc. ***All these technological factors are related to safety. They constitute kernel of Confiance.ai program and gather the main part of the activities***. Sociological factors, such as validation by peers, reputation in social networks, the attribution of a label by a trusted third party, etc. will complete the assessment of AI based system safety to improve their adoption.

Focuss on Security aspects

Security is here considered from the point of view of cybersecurity. It is a key dimension of trust can be included in safety consideration as attacks can trigger critical safety issues, but it is also often considered separately, for example for privacy aspects that not always triggers safety questions. Concerning attacks, if AI systems are, like all digital systems, susceptible to being attacked, hacked, compromised by "usual" methods (intrusion, decryption, virus, saturation, etc.), they have particular characteristics that make them particularly fragile to other types of more specific attacks. Adversarial attacks consist in injecting minor variations of the input data, during the inference phase, in order to significantly modify the system output. Since the famous example of the STOP sign not being recognized when tagged with labels, and the example of the panda being mistaken for a gibbon when a noise component is added, it is known that it is possible to compose an attack in such a way as to strongly modify the interpretation of the data made by a neural network. And this does not only concern images: one can conceive adversarial attacks on text, or on temporal signals (audio in particular, but also on any physical measures), etc. The consequences of such an attack can be dramatic, a bad interpretation of the input data can lead to a decision in the wrong direction (for example, accelerate instead of stopping, for a car). The report by NIST (NISTIR 2019) establishes an interesting taxonomy of attacks and corresponding defenses. In particular, it shows that attacks during the inference phase are not the only ones of concern. For instance, it is possible to pollute the learning bases with antagonistic examples, which naturally compromises the systems trained from these bases. If we add to this the "usual" security issues, as well as the multiple problems caused by deepfakes, it is clear that the AI security wall is now solid enough and close enough that it is essential to protect ourselves from it.

Even if first step of Confiance.ai if focused on safety, the issues of security are considered and will be subject of dedicated works in next phases.

Focuss on Interaction aspects

Interaction with humans can take various forms: speech, text, graphics, signs, etc. In any case it is not necessarily in the form of sentences. Interaction problems (in both directions) between AI systems and human operators and users can obviously cause safety issues if there is misunderstanding of critical situations. For example, if request made by users are ambiguous for the machine due to interaction problems, wrong interpretation of instructions can lead to undesired behavior (e.g. target a friendly vehicle, supply an inappropriate medication). The requests for proper interaction mechanisms in the case of autonomous vehicles have been well described in (Daimler et al., 2020), section 2.2.2.14 (quoting the introduction of this section): "Human-machine interaction (HMI) is considered a crucial element for the safe operation of SAE L3, L4 or L5 vehicles ... HMI should be carefully designed to consider the psychological and cognitive traits and states of human beings with the goal of optimizing the human's understanding of the task and situation and of reducing accidental misuse or incorrect operations".

The need for explanations of artificial intelligence systems is one of the measures of the regulation proposed by the European Commission (EC 2021), or of a draft standard concerning the certification of the development process of these systems (LNE 2021). *As they are key issues of safety and they will be considered by Confiance.ai in the next phases.*

Energy

The energy wall is well identified by some deep learning researchers. The seminal paper by Emma Strubell et al. (2019) found that training a large transforming natural language processing neural network, with optimization of the network architecture, consumed as much energy as five passenger cars over their lifetime (opposite). The paper by Thompson et al. (2020) went further, concluding that "the computational limits of deep learning will soon be constraining for a range of applications, making the achievement of important benchmark milestones impossible if current trajectories hold." This is a key issue in particular if we consider this subject more largely in terms of required or wished frugality of AI both in data, algorithms and computation resources. *As embedded systems are natural targets of Confiance.ai this subject will be considered through the angle of the impact of resources (energy, memory, data) optimization on the AI based system safety.*

(non-)Humanity

Finally, one have to mention a fifth wall is the one of the humanity of machines, or rather the one of their "inhumanity" (in the sense of "not being human"). It gather several hot subjects as: acquisition of common sense; causal reasoning; transition to System 2 thinking in the sense of Kahneman (2013). All components that we, humans, naturally possess and that artificial intelligence systems do not have. *Even if it is a crucial set of issues that could change completely the relation and safety of AI, it seems still to require long-term researches, and thus is not addressed by the program Confiance.ai.*

Overview of Confiance.ai approach

The program Confiance.ai is the technological pillar of the Grand Challenge "*Securing, certifying and enhancing the reliability of systems based on artificial intelligence*" launched by the Innovation Council. The two other pillars focus on standardization (norms, standards and regulation toward certification) and application evaluation.
Confiance.ai is the largest technological research program in the #AIforHumanity (2018) plan. It tackles the challenge of AI industrialization, as the very large-scale deployment of industrial systems integrating AI is a crucial stake for industrial and economic competitiveness. It has a strong ambition: breaking down the barriers associated with the industrialization of AI and equipping industrial players with methods adapted to their engineering. One originality of the program lies in its integrative strategy: it addresses the scientific challenges related to trustworthy AI and provides tangible solutions that can be applied in the real world and are ready for deployment in operations.
As defined by the European *commission (EC 2020), (EC 2021)* **trust** is the key objective for a deployment in respect to the European values. It can be defined through various points of views, details and encompass both engineering and usage aspects. Even if Confiance.ai has to consider all aspects, a particular effort is made to propose concrete and pragmatic answers for system and software engineering methods able to allow certification of AI based systems according to their criticality levels.

Confiance.ai has organized the program upon the four main stages of ML component development also identify by R. Ashmore, R. Calinescy and C. Paterson (2019): data management, model learning (or "design"), model verification and model deployment. The structure is completed by a transversal objective to define the methodologies for engineering and certification of AI based systems (Figure 1).

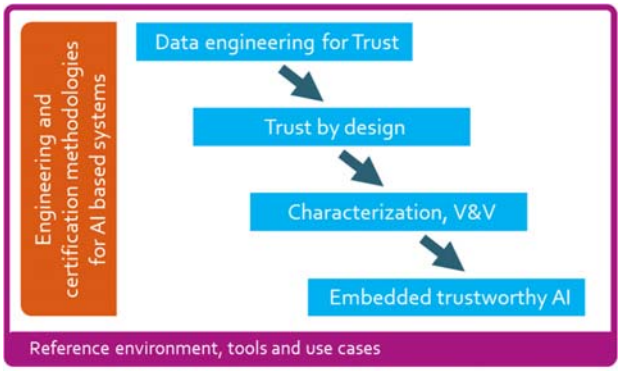

Figure 1: Confiance.ai program architecture

Each subject triggers several focused actions evaluated on use cases to help identifying and assessing the capacities of technologies to provide valuable arguments for safety assessment. The program adopts a strategy of progressive advancement: during the first year of the program, data-based AI solutions, mainly using neural networks, are the focus of research with application on image processing, time series and structured data. Then, in the following years, more complex problems and relevant industrial use cases will be looked at. Use cases using video, audio and text data will be added, as well as the introduction of other AI formalisms including knowledge-based and hybrid approaches. At the end of the program, the program will cover the whole spectrum of critical systems.

Technological and scientific challenges

More precisely, we identified more than 40 technological and scientific detailed challenges for the program. The list of challenges is subject to changes as we progress, it has already evolved since the launch of the program one year ago. The program adopted the term of "trust" to remain open to all possible factors ensuring an AI deployment that will be beneficial for humans. In practice, at least for the first phases of Confiance.ai, "*Trustworthy*" could be understood as "*Safe*" as the focus is this of ensuring, evaluating, certifying the AI based system safety. As of now, the challenges belong to three main categories and eight subcategories:

1. Trustworthy system engineering with AI components
- Qualify AI-based components and systems
- Building AI components with controlled trust
- Embeddability of trustworthy AI

2. Trust and learning data
- Qualify data/knowledge for learning
- Building data/knowledge to increase confidence in learning

3. Trust and human interaction
- Trust-generating interaction between users and AI-based system
- Trust-generating interaction between designer/certifiers and AI-based systems

The first category gathers all aspects of designing and evaluating AI components for trust (safety). Issues such as performance, robustness, verification, proof, monitoring and supervision, as well as hybrid systems mixing data-based and knowledge-based solutions, belong to this category. Since the major application area of the program is critical systems, we also put an emphasis on embedded AI, aiming at maintaining the desired properties in environments where memory, computation capacity, energy usage and real time behavior are constrained.

The second category deals with data and knowledge. Here we consider subjects such as data preparation, data augmentation (when the available data are not sufficient), heterogeneity of data, domain adaptation, mixing data-based and knowledge-based models. Another key consideration if that of the ODD (operational design domain) in which an automated function or system is designed to properly operate.

The third category puts the emphasis on proper interaction between humans and AI-based systems, focusing on three types of interaction: (i) during the design phase; (ii) for certification by authorities; (iii) when in the hands of final users with major issues being transparency and explainability.

To make things more concrete, let us take two examples of detailed challenges: (i) in the first category, we aim to develop components integrating self-monitoring of staying within the ODD boundaries. For this purpose, we need a clear definition of the ODD, as formal as possible; alert mechanisms when the system approaches the boundaries; and stopping mechanisms when the system has exited the ODD. (ii) In the third category, we look at methods of explanation corresponding to the needs of users, designers and certifiers. There is a variety of explanation methods for different kinds of problems and data (e.g. saliency maps for images, logic-based explanations for numerical data, text-based explanations for knowledge-based approaches etc.); we analyzed and tested a dozen available explanation methods and tools, but at this time none of them brings a full solution to the question, more research is needed.

A validation strategy based on industrial use cases

Confiance.ai is an industry-oriented project. Its outputs are expected to be usable by industrial partners within their software engineering process. A way to achieve this objective is to validate the produced methods and tools on industrial use cases.
Use cases are formally defined by

- A feature implemented with AI-based technologies.
- An acceptability issue raised by any kind of authorities.
- Access to the data or the knowledge base used by the feature
- Involvement of the feature product owner himself for the evaluation of the proposed methods and tools

To reach this goal, the project must perfectly understand the arguments that will convince the validation authorities. That is the reason why the involvement of the product owner is crucial. Each tool provided by the project should be a step towards the demonstration of the AI-based system safety. Furthermore, because this demonstration will rely on the way the function has been developed and validated, the use case carrier must be transparent about the way he generated the function: development process, source code, training and validation data base, validation process…

Providing a use case to Confiance.ai is thus not that simple. It is sometimes difficult to share data or knowledge without sharing intellectual property or confidential information. A part of competitive advantage could be in selected network architecture. These aspects can be circumvented by providing representative publicly available data or well-known networks instead of the real artefact used by the industrial partner. But in this case, these public use cases come rarely with all the information on the development context, in particular regarding feature related to the quality process, and consequently can be used only partially to assess the tool results on the AI function, and less for evaluating the soundness of methodological proposals at system level.
As being developed in a research project, the AI-based features are often under development or at POC status, their integration in critical industrial systems is not expected at short term when plenty of other critical issues are to be managed today. The connection between safety system requirements and the software technical proposed solution at the component level will be the major challenge of the project.

Nevertheless, a first set of use cases have been proposed by Confiance.ai partners. They are all supported by data-based AI (implemented through artificial neural networks technologies) dealing with vision, time series and surrogate models.

2D vision		Visual Inspection		Surrogate
Road scene understanding	Classification in Aerial pictures	Welding quality	Indication detection	Look-up table (ACAS XU)
Valeo	Airbus	Renault	Safran	Airbus

Other use cases will be integrated for the second year to complete the panel of the AI challenges, for example concerning: Natural Language Processing and Audio processing.

To illustrate the context and process of work around the use case, we shortly describe here the "Welding" use case, by Renault. The implemented feature is a vision- based detector of the quality of a welding. This feature is expected to assist the human operator in tracking the possible default on welding point. This feature has been implemented with neural networks techniques because they allow a simple learning phase doable by non-software experts. These welding points, on the chassis, are involved in the safety of the vehicle, their control is critical. Despite the very good performance of the classification, the quality management of the factory does not trust the efficiency of the AI based system. This is a hard issue of acceptability. The objective in Confiance.ai project is to build justification arguments in order make this feature accepted by the quality management.

First results on a representative use case

After less than 1 year (project starts in 2021), some tools have been evaluated on the selected industrial use cases. For instance, we have developed several ways to evaluate the robustness of a classifier. One of it, illustrated on Figure 3, is to add noise to the input pictures (lightning conditions, gaussian blur, motion blur, dead columns, dead pixels…) and check the evolution of the classification accuracy.

Figure 2: Image perturbation examples for robustness evaluation

Based on an original welding picture (middle), sensor troubles have been simulated (dead pixels on the left, loss of focus on the right). The graphics below (represent the evolution of the error according to the amplitude of the noise for several pictures. This very simple example illustrates the necessary connection with the use case owner: what kind of noise is relevant? Which noise amplitude is realistic? How to fit such a robustness evaluation with the quality requirement?

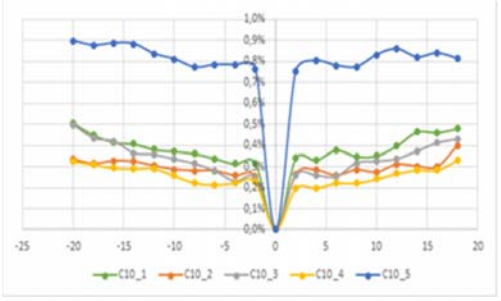

Figure 3: Example of tool output evaluating accuracy variation depending on different brightness variations.

However, explainability is an important aspect to reach the acceptability of AI. We have evaluated several existing methods: Rise (Petsiuk & al 2018), Lime (Ribeiro & al 2016), Occlusion (Zeiler and Fergus 2014), KernelSHAP (Lundberg and Lee 2017)… The methods proposed within the Xplique and GemsAI libraries (developed by ANITI and the DEEL project) have been used to highlight the parts of the picture that have been used to take the decision about the quality of the welding. Figure 4 demonstrates that the AI system pays particular attention at a certain part of the welding to classify it.

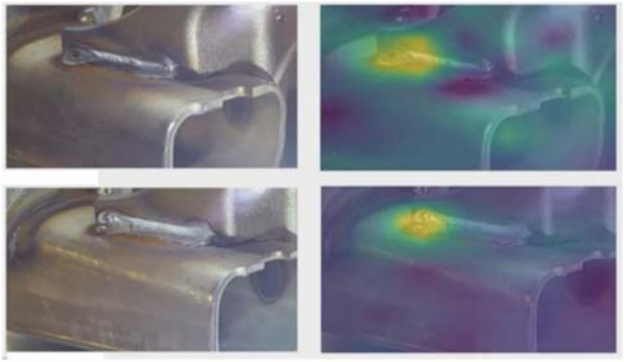

Figure 4: Explainability of classification decision

The output of explainability can be used by the software developer, to validate the good behavior of the developed model. But it can also be used by the person in charge of quality check, on the manufacturing line. At last, it can be used to convince the quality manager that the AI is trustworthy because it takes its decision based on the right observation.

Many other tools have been developed to characterize and monitor the behavior of AI based components. We also propose methods to improve the robustness of neural networks: 1 Lipschitz network (Tzuzuku & al 2018), randomized smoothing (Cohen & al 2019), adversarial training (Bal-

unovic and Vechev 2019)… but also, for instance, to validate the quality and the completeness of the data used for training : Pixano developed by CEA (Dupont 2020), Debiai by IRT SystemX.

The black box cases

Even if AI components are developed internally by the industrial partners, some others will be bought off-the-shelves. For instance, the automotive industry uses smart cameras, developed by other companies. These cameras integrate AI-based features for which source-code, training data base, development methods are not accessible. Such use cases are to be considered as well by the project that will develop tools and methods to evaluate, validate and monitor such features without the requiring the 4 criteria exposed before.

In this case, what is required is of course the access to the device but, mainly, the clear statement of the product owner's expectations. What should be demonstrated? Which kind of validation could be decisive for the owner? In this case, the output of the project will be more the good questions to ask to the supplier than technical tools to answer these questions.

The confidential use cases

For some reasons, mentioned above, partners will not be able to share their use cases. Anyway, they want to validate the methods and tools proposed by the project.

As a matter of fact, it represents another way for the project to validate it outputs. Instead of providing a "certified" use case, as it does for selected use cases, the project will provide methods and tools to be used to "qualify/certify" a use case. Each partner can use these methods and tools either by using the whole environment provided by the project or by integrating them in its own development cycle and will validate, internally, the efficiency of the provided outputs. If the "qualification/certification" is doable internally, the project is successful: the aim of the project is not to provide "certified" use cases but methods and tools to "certify". If the provided methods and tools are not good enough, it will be a very rich feedback to improve them within the project.

In a way, non-sharable use cases will almost be more useful than sharable ones because they will demonstrate the relevance of the Confiance.ai project, able to deal with use cases it was not specifically designed for.

Conclusion

Confiance.ai is the largest French project on AI focusing on trust, with particular concerns on safety critical applications at different levels of criticality. It targets setting up a complete tool chain for the development of trustworthy AI based systems. For that Confiance.ai encompasses the whole cycle with the focus of ensuring trust at each stage, from data management, AI design and AI validation to deployment. This includes the system qualification by defining the element required for qualification accord to the requirements of respective applications domains (aeronautics, automotive, defense, energy…).
Working process is iterative and incremental and strongly attached to real operational industrial use cases on which all the different tools and methods (either for existing ones and for those developed in Confiance.ai) are evaluated. Focus has been made for the first year on neural network -based AI for applications requiring real qualification but with low criticality (for example with human remaining in the loop).
First results shows that mathematical approaches for robustness or explainability could provide interesting elements to ease the qualification. Next steps will be completing the chain, for example by addressing the question of ODD definition and management and with integrating applications using hybrid AI with the objective to obtain within the 4 years of the project both methodological guidelines and tool chains adapted to each of the partners engineering contexts.

References

AI4Humanity (2018) https://www.aiforhumanity.fr/

Ashmore, R., Calinescu, R., and Paterson, C. (2019). Assuring the machine learning lifecycle: Desiderata, methods, and challenges.

Bengio et al (2021) Yoshua Bengio, Yann Lecun, Geoffrey Hinton. Deep Learning for AI. Communications of the ACM, July 2021, Vol. 64 No. 7, Pages 58-65

Balunovic, M., & Vechev, M. (2019, September). Adversarial training and provable defenses: Bridging the gap. In International Conference on Learning Representations.

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019, May). Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning (pp. 1310-1320). PMLR.

Darpa (2017) https://www.darpa.mil/program/explainable-artificial-intelligence

Daimler et al. (2020), Safety first for autonomous driving, https://www.daimler.com/innovation/case/autonomous/safety-first-for-automated-driving-2.html

Dupont, C., Ouakrim, Y., Pham, Q. C. (2020) UCP-Net: Unstructured Contour Points for Instance Segmentation. IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020

EC (2019) https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_policy_and_investment_recommendations.pdf

EC (2020) https://ec.europa.eu/.../commission-white-paper-artificial-intelligence-feb2020_en.pdf

EC (2021) https://ec.europa.eu/france/news/20210421 nouvelles_regles_europeennes_intelligence_artificielle_fr

GemsAI : https://github.com/XAI-ANITI/ethik

Hidalgo C. (2021) https://www.judgingmachines.com/

ISO (2020a) https://www.iso.org/obp/ui#iso:std:iso-iec:tr:24028:ed-1:v1:en

ISO (2020b)
Information technology — Artificial intelligence — Risk management - ISO/IEC CD 23894

Kahneman D. (2013) Thinking, Fast and Slow, Farrar, Straus and Giroux;

Littman et al. (2021) Michael L. Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfield, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, and Toby Walsh. “Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report.” Stanford University, Stanford, CA, September 2021. Doc: http://ai100.stanford.edu/2021-report.

LNE (2021) Laboratoire National de Métrologie et d’Essais; Processus de conception, de développement, d’évaluation et de maintien en conditions opérationnelles des intelligences artificielles

Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).

NISTIR draft 8269 (2019), A Taxonomy and Terminology of Adversarial Machine Learning, E. Tabassi et al, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Russell S. (2020), Human compatible: Artificial intelligence and the problem of control
Stuart Russell, Penguin Books, 2020

Strubell et al (2019) E. Strubell, A. Ganesh, A. McCallum; Energy and Policy Considerations for Deep Learning in NL , https://arxiv.org/abs/1906.02243v1

Thompson N. et al. (2020) The Computational Limits of Deep Learning, arXiv:2007.05558v1

Tsuzuku, Y., Sato, I., & Sugiyama, M. (2018). Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. arXiv preprint arXiv:1802.04034.

Xplique : https://github.com/deel-ai/xplique

Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.