SIMPLY RATIONAL

EU AI Act – An Opportunity!

Dr. Niklas Keller, Prof. Dr. Florian Artinger, Dr. Malte Petersen, & Prof. Dr. Gerd Gigerenzer

Simply Rational – A Max-Planck Spin-Off

There is a good reason why the EU wants to regulate the use of artificial intelligence (AI). In recent years, high-profile cases have been uncovered in which artificial intelligence systems failed dismally to the detriment of both citizens and businesses. In its current state, the EU AI Act confronts businesses with significant regulatory uncertainty—first regarding its system for risk-classification and the question of in which category a businesses' AI systems will end up in, and second regarding the resultant regulatory requirements that will be placed upon these systems. But the requirement of greater transparency for high-risk systems in the EU AI Act also provides an opportunity. As it turns out, research at Max-Planck-Institutes and beyond has shown that simple and transparent algorithms based on human information processing principles ("Psychological AI") can often outperform even the most modern and complex black-box algorithms from machine learning. By implementing more transparent AI, businesses can significantly lower the regulatory uncertainty surrounding the risk classification section of the AI Act, as many of the regulatory hurdles and costs for implementing pre-market assessment and post-market monitoring can be radically reduced through the use of transparent algorithms.

Introduction

The pressure is on. The German Antitrust Authorities recently asked the 50 largest mail-order companies and all the larger credit information bureaus to lay open the inner workings of their algorithms. Even if the form which the EU AI-Act will ultimately take is not clear yet, regulators already now demand more transparency, test boundaries, create precedents, and try out different ways of regulating AI once the AI-Act will be fully rolled out. Sticking one's head in the sand is no longer an option. The time to get prepared for the new regulatory reality is now.

The AI-Act didn't drop out of the blue sky

While evil tongues may whisper that the EU AI-Act is simply another power grab by EU bureaucrats, it is also worth remembering that the AI-Act comes in the wake of an increasing number of AI deployments that have retrospectively turned out to be potentially catastrophic to the wellbeing of hundreds of thousands of citizens. Some high-profile cases include:

> Judicial Decisions: The "Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)" algorithm has supported millions of bailing and other judicial decisions in the US. A study, however, showed that COMPAS, using over 130 variables, predicts recidivism no better than ordinary people with no legal background using only 7 variables. In addition, the black-box nature of COMPAS makes it difficult to determine whether it discriminates against minorities.

Facial Recognition: Google Photos' image recognition software classified two African-Americans as "gorillas". Similarly, Amazon's facial recognition AI had a far higher accuracy for males (3% error rate) than for females (19% error rate) and the worst accuracy for black females (39% error rate).

Oncology: IBM's state-of-the-art AI "Watson" was to revolutionize oncology: improve the care of cancer patients, recommend the best treatments, and completely overhaul the search for new pharmaceuticals. But it was never able to deliver on the promises of IBMs marketing department, endangered the lives of patients, and was essentially sold for parts in 2021, including the patient data. This was after it had already been used in trials to support the care of thousands of patients.

These are just a small selection of high-profile cases. Particularly in medicine, a sector which due to its size has been receiving the brunt of AI-initiatives, reports of "highly sophisticated" AI-systems failing miserably in their intended tasks are emerging on an almost daily basis. Together, these systems have already "supported" healthcare decisions in millions of cases with untold consequences for patients and doctors alike. Regulators are therefore rightly concerned about the performance of AI algorithms in sensitive areas such as healthcare, law, or finance.

The great unknown: what risk-level an AI will have

The way that the AI-Act categorizes risks (see Figure 1) is in need of improvement. While the section on risk classification has been significantly improved compared to first drafts, it is still the case that sometimes specific technologies are categorized on their own without mention of their goal or purpose (e.g., AI chatbots are all still considered low risk). Sometimes goals are categorized on their own without any mention of the underlying technology (e.g., all forms of social scoring). Sometimes goals and technologies are lumped together (e.g., real-time, remote biometric identification systems used in public spaces for law enforcement). And most of these categories are not mutually exclusive. Is a (non-manipulative) AI chatbot that causes people interacting with it to commit suicide more frequently a minimal risk system? Is an eBay or Uber rating used to verify the trustworthiness of a seller or driver not a form of social scoring even though it reduces risk for the people using the services?

Figure 1. The three risk categories of the EU AI-Act and according categorization of AI technologies or areas of applications.

Furthermore, the list of High-Risk Technologies has been expanded throughout iterations of the AI-Act and the criteria the Act proposes for adding additional sectors to this list remain open to significant degrees of interpretability. They may, in fact, likely present one of the major battlegrounds of lobbying efforts after introduction of the Act. This is reflected in the evaluation of the Act by the Council of the European Union from the 6th of December 2022, which highlights continued uncertainty regarding the definition of AI and suggests changes especially to the risk classification system (both the list of high-risk technologies and the ways of defining these). Until its final iteration (which in any case will continue to be able to be edited by adding more sectors to the list of high-risk technologies), we can continue to expect significant changes. As a consequence, uncertainty about which risk category a companies' product or service will end up in will remain high for the foreseeable future.

The Complexity Illusion

When we look at the reasons behind the AI-Act, "trust and transparency" are certainly the most frequently mentioned by Brussels. Even with all the initial hype surrounding it, the many recorded failings of AI have understandably eroded the trust. Need for greater oversight is the consequence. Where there is no trust, there must be at least transparency. But why is so much of the AI that we deploy today so intransparent? One reason is a pervasive "complexity illusion" in business, politics and society at large.

The assumption that more information and more sophisticated, complex models with greater flexibility (i.e., more free parameters) automatically outperform simpler models is so deeply entrenched that it is rarely discussed and almost never challenged. Simpler approaches may be used to perform certain tasks, but only because the effort required to apply a complex model, or the additional information or computational power needed, simply outweigh the benefits of increased performance. This is termed the "effort-accuracy trade-off". The more effort you put in, the better the accuracy (with diminishing returns). This assumption also directly translates into the (incorrect) claim that using easy-to-understand, simple, and transparent algorithms always reduce prediction accuracy. It (mis-)led the Defense Advanced Research Projects Agency (DARPA), a military research agency and one of the largest research-grant providers in the United States, to publish this figure in 2016:

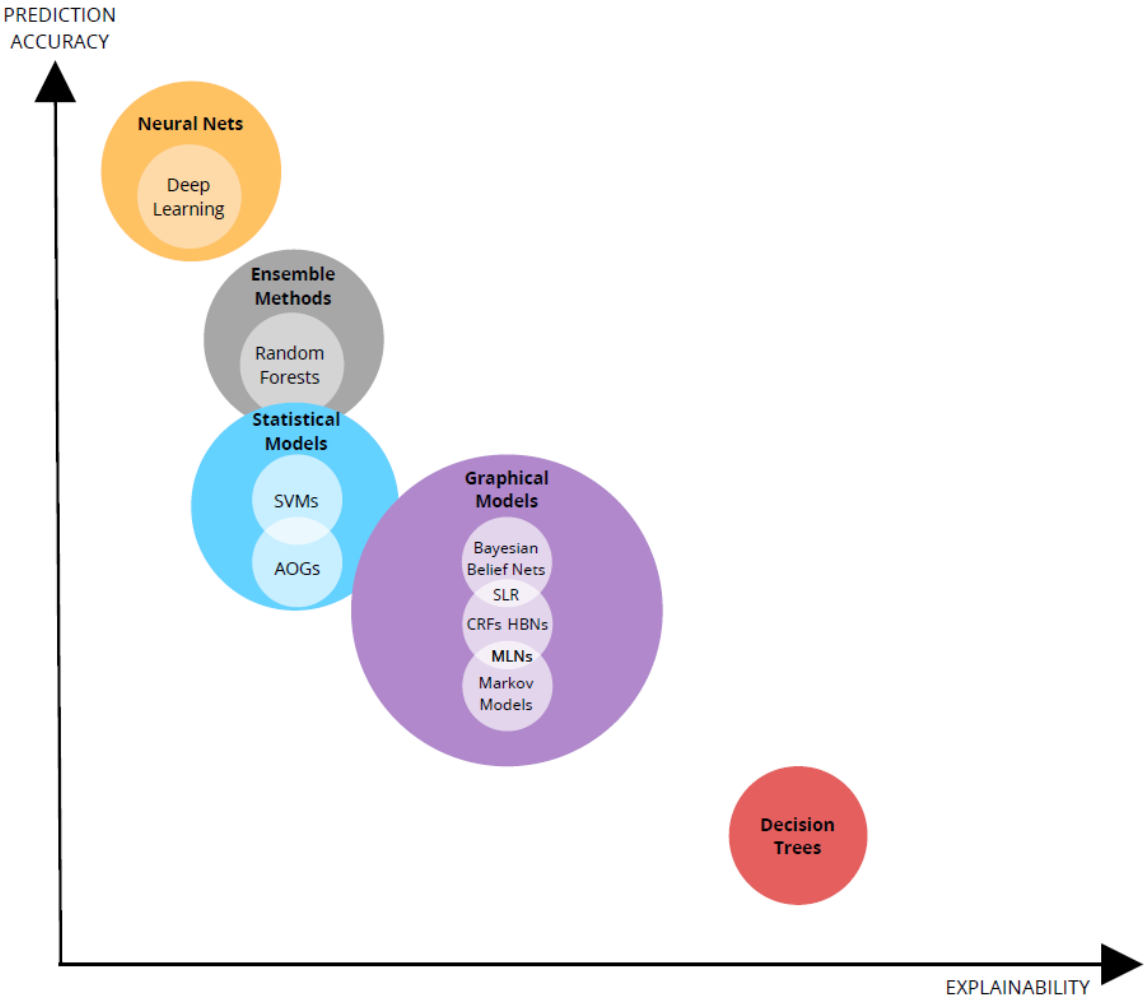

Figure 2 (from DARPA 2016). The Effort-Accuracy Trade-Off as it translates into the Explainability-Accuracy Trade-Off. The (faulty) assumption here is that models of lower complexity (e.g., decision trees at the bottom-right) have higher explainability, but lower prediction accuracy.

At the top-left corner, we have the pinnacle of complexity in modern artificial intelligence: the neural network, where deep learning happens. Too bad, that no one really understands the inner workings of such models once they have been trained. Not even the engineers that built these models! When Google Photos classified African Americans as gorillas, there was no line of code that the Google engineers could swap out or delete. The only way they could solve the problem was by removing the entire category of "Great Apes" from the possible choices.

Scrounging about in the lower right corner: the lowly decision tree—easy to understand and explain, but allegedly with terrible performance. Does the EU AI-Act condemn us to use the worst technologies for precisely the most important aspects of our lives, business, and societies?

Transparent Algorithms can outperform black-boxes

Luckily, no. It turns out, the Effort-Accuracy Trade-Off does not generally hold true. The idea that more complex, sophisticated models always outperform simpler ones is demonstrably false. Models that are more simple, more transparent, use less data, and combine this data in less complex ways can often

outperform highly sophisticated and complex AI black-box models. Less can be more! Here are three illustrative examples.

Judicial Decisions: In the wake of the COMPAS study, our colleagues from Microsoft Research and Stanford University developed a transparent algorithm based on principles from psychology and cognitive science (what we term "Psychological AI") and compared it to the performance of a complex AI model (a so-called random forest). Both methods had all of the data at their disposal. The random forest ended up integrating all of the 64 available variables in highly intransparent ways. The transparent algorithm only used two variables (age and the number of previous times that the defendant did not appear before court) and integrated these into a transparent scoring method. Despite these vast differences in use of data and algorithmic complexity, the algorithms performed about the same (and both significantly better than the judges).

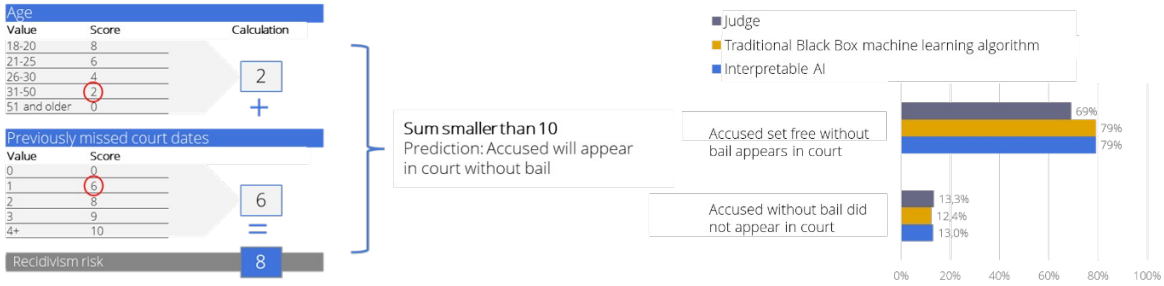

Figure 3. A simple, transparent algorithm to predict recidivism. The algorithm (left) only takes only two factors into account (age and number of previous missed appearances in court) and integrates these using a simple scoring method. Its performance (right) is better than that of experienced judges and on par with that of a complex random forest.

Predicting Pandemics: In 2008, Google Flu Trends was to revolutionize (as all things AI do) the prediction and tracking of flu pandemics. The idea was simple: rather than using the slow process of the U.S. Centre for Disease Control collecting data on flu diagnoses from doctors, Google's AI would analyze search terms for flu and flu symptoms to predict if a flu was spreading in a particular area. Initially, 50 million search terms and over 100 million different prediction models were evaluated. The engineers then selected 45 (secret) search terms, but after a series of failures to predict flu pandemics (swine-flu, MERS, SARS 1) the Google engineers increased this to around 160 search terms. 2015, the project was quietly shelved. Instead of more complexity, a team including researchers from Simply Rational went for the opposite approach: what if under high uncertainty, it is best to ignore the past and rely on only the most recent data points? The model we developed used a single "smart" data point, the number of flu-related doctors' visits of the current week, and predicted the same visits for the next week. Across eight years of predicting the flu (2007-2015), this smart rule had an error rate of about *half* (0.2% vs 0.38%) that of Google Flu Trends. Less is more.

Customer Base Analysis: Knowing which customers to target (for example for marketing campaigns) and which of the existing customers are and are not likely to make purchases in the future is important for improving the efficiency and success of a business. Already in 2008, two former colleagues showed that a simple managerial rule of thumb can beat a mathematical "optimizing model" from marketing (Pareto/NBD-Model) across three markets (see fig. 4, left). The simple model, called "hiatus heuristic", is used by sales managers across industries: "If a customer has not made a purchase for X months, classify

them as inactive, otherwise as active." We recently teamed up to extend this analysis to 61 (retail- and non-retail) datasets, including logistic regressions and again random forests in our comparisons. Again, the transparent heuristic algorithm could not only hold its own against these far more complex competitors, but actually substantially outperformed them across all of these datasets and applications (see fig. 4, right).

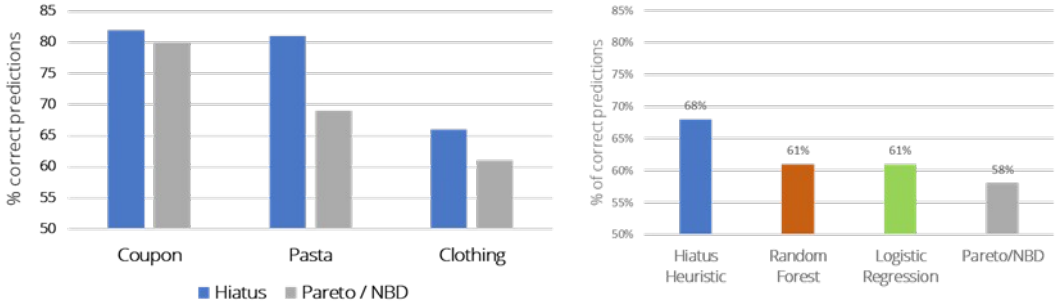

Figure 4. Predicting customer activity. A transparent heuristic not only beats a complex mathematical model from marketing (left side: Hiatus Heuristic; blue / Pareto/NBD; grey), but also statistical and modern machine learning methods across 61 applications (right side: hiatus; blue / random forest; red / logistic regression; green / Pareto/NBD; grey)

Transparency: In an uncertain world, less can be more

These examples are just the tip of the iceberg of studies and applications that have shown that simple models can make better predictions *even if a lot of data is available*. The reason is what is called the "bias-variance trade-off". Any prediction algorithm can produce two types of error: bias, which is a systematic deviation of the average estimated value from the true value, and variance, which is a spread around the average estimated value. If an algorithm is too simple, it runs the risk of "overlooking" important factors and will be systematically and predictably off – resulting in a high bias. If an algorithm is too complex, however, it risks incorporating irrelevant factors into its predictions, increasing estimation error, and resulting in high variance. For many complex algorithms applied to real-world problems, the prediction error from variance can be far greater than that from bias.

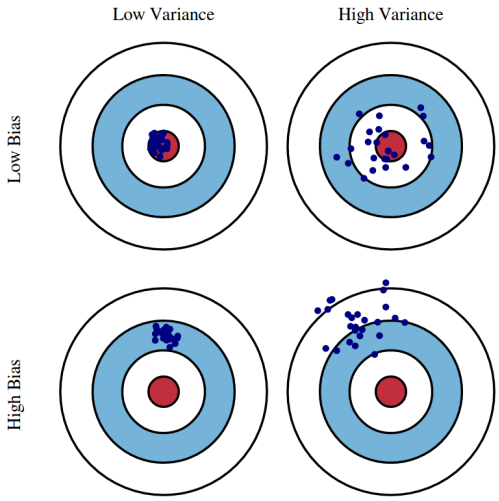

Figure 5. A bias (the bottom two targets show high bias) is a systematic deviation of the average estimated value from the true value, the bull's eye – think of a shotgun with a skewed barrel always missing the target by the same predictable degree. Variance (the right two targets show high variance) is

the spread of the individual values (the bullet holes) around their average – think of a sawed-off shotgun having a bigger spread than with a longer barrel.

Bias and variance are intricately connected to the environment in which they are applied. If an environment is stable and the past can be expected to be like the future, a complex, sophisticated AI algorithm using all available data will be able to tease out ever smaller nuances to arrive at ever better predictions. If the world is not stable, however, its many free parameters present "points of attack" for uncertainty where estimation errors result in increased variability and fluctuations of predictions and a performance drop. In contrast, a simple, transparent algorithm, that relies on only a few robust factors for its predictions, has fewer points of attacks. As a consequence, even if a simple Psychological AI and a complex black-box AI use the same three robust factors to make predictions, but the complex AI uses an additional 30 other factors for its predictions, it is likely that the simple AI using fewer factors will outperform the complex one, because it is not led astray by useless information.

What does the EU AI-Act mean for business? An opportunity!

Knowing that the transparency of algorithms does not need to stand in conflict with their prediction accuracy can significantly impact how one approaches the AI-Act and the regulatory uncertainty surrounding it. A major uncertainty concerns the section on risk stratification, i.e., it is hard to tell for most organizations using AI to drive their business which of their algorithms will end up in the high-risk category. But there is less uncertainty surrounding what the AI-Act will require organizations to do in order to show that they can effectively manage these risks. Currently included are:

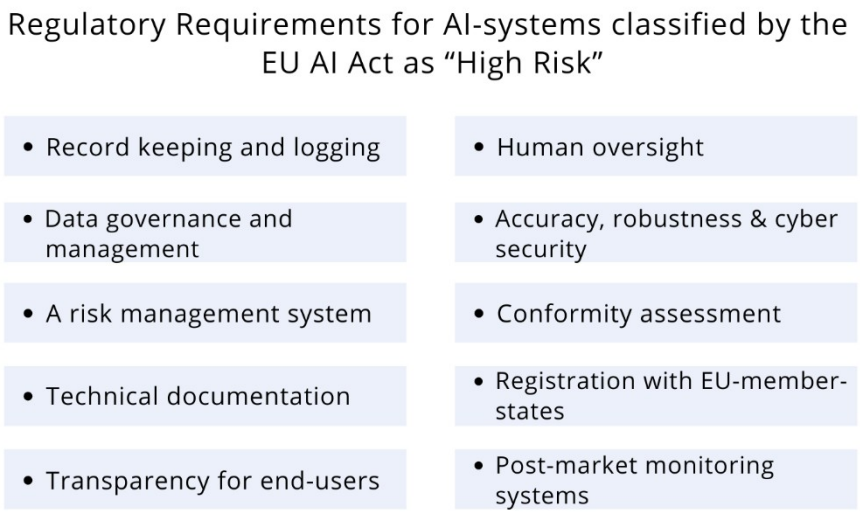

Figure 6. Regulatory measures required for high-risk AI. Many of the requirements are far more easily addressed when deploying transparent algorithms rather than black-boxes.

In the final AI-Act, one or two further measures may be added, one or two may be dropped, or they may be structured in a particular way. But overall, there won't be huge surprises in terms of the systems and processes that regulators will require organizations to have in place to assure compliance with the AI-Act. A glance at each of these requirements reveals that transparent Psychological AI facilitates many aspects both pre-market assessment and post-market monitoring and can positively impact the cost-benefit calculation for many of the above regulatory requirements. Record keeping, data governance & management, and technical documentation are easier. Risks can be more pro-actively managed and risk

managers or compliance officers do not need IT-degrees to understand what is happening, including identifying risk of biases and other ethical or moral breaches. Human oversight and transparency are the default and accompanying conformity assessments less effortful for all entities involved. At the same time, as discussed above, the accuracy of simple models will likely in many cases be on par with or even higher than current, complex solutions. And their robustness, i.e., applicability to new cases outside of those the algorithm was trained on will likely be higher.

In addition to addressing most of the requirements that the AI-Act is likely to impose, psychological AI algorithms also have advantages from a business and performance perspective. They allow easier integration with human knowledge & expertise. This leads to better overall system performance. An example is chess, where the best AI beats the best humans, but when both team-up together, they can beat the best AI. In many situations, transparent Psychological AI algorithms can also be implemented, maintained & monitored more cheaply and easily. Whenever it is important to keep humans in the loop, both the utility and the output of a system using a Psychological AI can be more easily communicated, justified, or defended. This also ties into another major advantage of transparency: in any field such as insurance, credit assessment, health, jurisprudence, etc. where a behavioral change of the customer can positively impact a business, a Psychological AI that is based on factors that customers can actively influence serves as guidance for that behavior in a process of truly shared decision making between company representatives and existing or potential future customers. When it comes to what kind of AI your business should employ, ask yourself these three questions:

Does the complex, black-box AI algorithm outperform the transparent and intuitive Psychological AI?

→If not, use Psychological AI.

If yes: Does this performance advantage translate into better system performance (i.e., the performance of the human and the AI together)?

→If not, use Psychological AI

If yes: Is this performance advantage large enough to trade-off against the more effortful and costly compliance with regulatory requirements for black-box AI of the AI-Act?

→If not, use Psychological AI

Only if you can definitively say YES to all, should your business consider black-box AI models.

The AI-Act: A Chance for Progress

The complexity illusion (see Figure 2), the unquestioned and widespread assumption that complex algorithms using more data are always better than simpler ones, is a primary driver of the increasing intransparency of AI tools and subsequent lack of trust we experience today. We can do better. We know that complex, intransparent AI models work better in stable environments and not-so-well in instable environments. And we know that the transparent Psychological AI models discovered over the last 20+ years work better in unstable, dynamic environments. With this in mind, we make the following calls for the AI-Act:

> We suggest the AI-Act should include provisions setting simple, transparent algorithms as benchmark. Only if a complex algorithm can demonstrate definitive and meaningful performance advantages over a simple method, should it be it be deployed. Arguments based on national security or trade secrets should be heavily scrutinized by an independent specialist EU body.

Any algorithm whose output has the potential to directly impact the life-situation of an individual must be transparent, and not only to regulatory bodies but also to the general public. Such transparency is preferably achieved by using transparent Psychological AI models in the first place, rather than by post-hoc methods (such as Shapley values or LIME), which are themselves only approximations to the underlying black-box algorithms. The same goes for simulators for intransparent models that only use part of the factors of the underlying black-box models.

Arguments that end-users and other stakeholders are incapable of understanding such transparency must be empirically supported. Even if shown that full transparency cannot be understood by most stakeholders, different levels of transparency should be deployed for stakeholders with different levels of technical competence.

We have made our lives far more complex and difficult than they really need to be. There are many situations in which simple, transparent algorithms outperform the intransparent black-box solutions that are currently used by businesses. In this light, the EU AI-Act can be viewed as an opportunity for organizations to reflect their use of black-boxes for making decisions.