Validate AI

Our position on tackling AI risks.

12th April 2024 (V3)

Co-Authors

This paper represents the ***Validate AI Steering Group*** position and approach to tackling AI risks, with thanks to:

Dr A. Bellotti, School of Computer Science, University of Nottingham, Ningbo
Dr Zeynep Engin, Data for Policy CIC
Professor D.J. Hand, Department of Mathematics, Imperial College, London
Seb Hargreaves, The Operational Research Society
Gile Herdale, Herdale Digital Consulting
Ed Humpherson, Office for Statistical Regulation
Dr Mark Kennedy, Data Science Institute, Imperial College, London
Shakeel Khan, Validate AI CIC
Professor Marta Kwiatkowska, The University of Oxford
Gilbert Owusu, Director of Data & AI at BT

1. Context

Artificial intelligence (AI) comprises an expanding family of software systems that are able to undertake tasks that would traditionally require human intelligence. Examples of these are game playing AI such as AlphaGo, recommender systems, medical diagnosis systems, scheduling software, automatic translation systems, the software for self-driving cars, generative AI, and chatbots.

From the beginning of 2023, we have seen a radical development of chatbot technology using large language models (LLM) based on sophisticated deep neural network technology and, in particular, generative pre-trained transformers (GPT). In particular, the release of OpenAI's ChatGPT in November 2022 has created great interest. ChatGPT and its next generation GPT4 are able to perform a wide variety of tasks that go beyond those their designers originally intended, such as writing poetry in different styles on any given subject, solving mathematical problems, conducting research and writing essays, programming from English language descriptions, achieving pass marks in examinations, and much more.

However, while the capabilities of ChatGPT and the potential it hints at for future generations of AI software have created much excitement, it has also led to anxiety within the AI community, the media, governments and the wider public, to the extent that the risks of AI as a disruptive technology are now a matter of public debate and concern.

The potential risks are various and can be summarized as:-

- **Risk of poor performance** and degradation of performance over time.
- **Data risk** associated with poor quality data, dark data leading to bias or discrimination, improper use of data, or poor data security.
- **Misaligned AI objectives**: the AI system may not complete the task as intended, or may perform its task in a way which is unanticipated and inappropriate.
- **Fragility:** the AI system may not be robust and may behave in strange ways in unexpected scenarios.
- **Misuse:** with poor controls, AI systems could be misused by the organisations that built them, or used by external users for fraudulent activity such as identity theft, or other criminal activities.
- **Human/AI interaction** may introduce harm to humans, by analogy to the way that social media can adversely affect human behaviour. This may manifest as misinformation with detrimental psychological effects and adverse outcomes.
- **Concerns about AI in Society:** such as the effect of AI on human employment in certain sectors and more generally, disruption to the economy, or risk of singularity.

In many organisations there is now a tension between those who see an urgent need to press ahead with AI development in order to achieve competitive advantage, provide better customer experience, or contribute to social benefit, and those who are concerned about the risks of deploying AI and the possibility of harm to humans and bias against protected groups.

2. Position

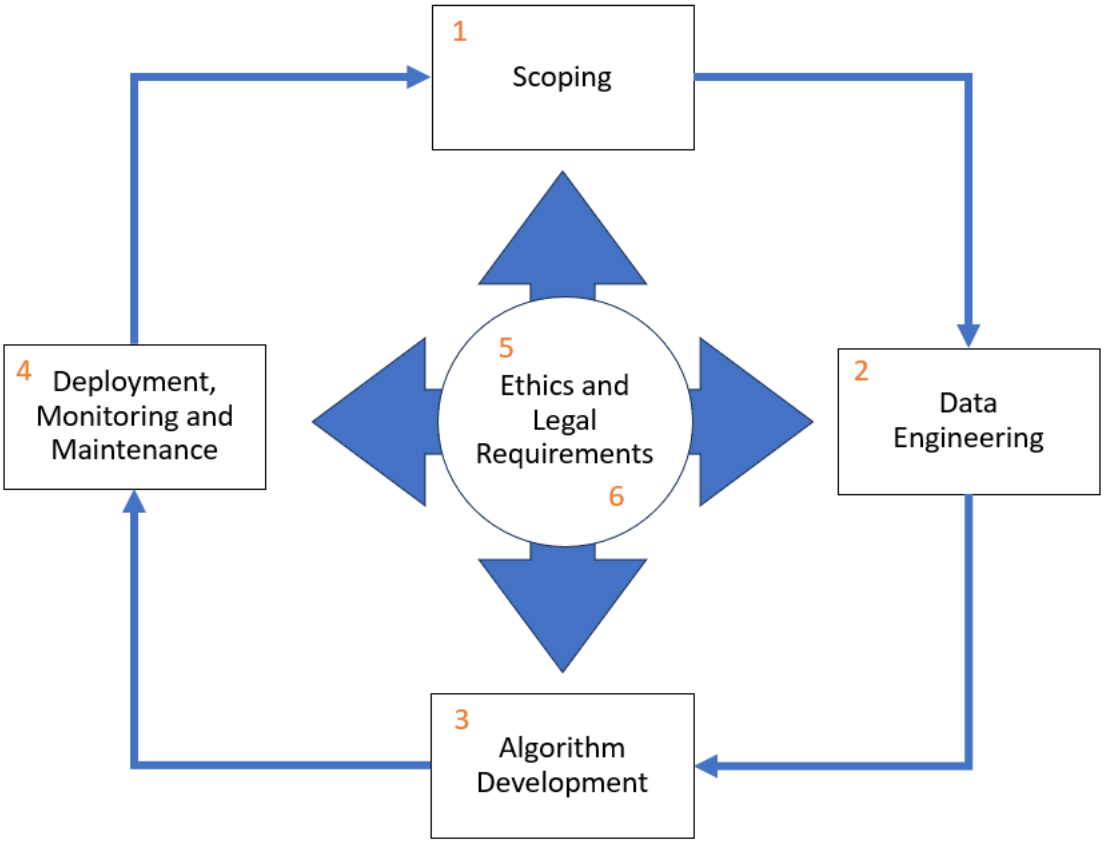
AI has huge potential to benefit us all, improving quality of life in contexts ranging from medical care to financial services. For this reason, we support the development of AI systems, but only when this is done in a way that takes comprehensive account of potential risks.

We believe AI systems need to adhere to the three tenets of being ***fit for purpose***, ***ethical***, and ***technically sound*** to be trusted.

Validate AI CIC is an organisation that has been conscious of the significant risks of AI since 2018 and has been championing discussion around responsible and trusted AI, promoting a ***practitioner-centric*** code of practice. This can only be achieved by inclusion of a diverse group of users and experts from the community convening to achieve consensus and to ensure the tenets of trusted AI are met.

To evidence AI trustworthiness, we prescribe the following six pillar approach which has been developed in consultation with key stakeholders including developers and customers of AI systems. This framework is practically applied as an assurance process via detailed checklists for each of these factors and represented in the diagram below.

The Six Factors of AI Assurance.



1. **Scoping.** Clarifying the objectives, mapping out a solution, and exploring whether AI is the right approach.

2. **Data Engineering**. Biased, incomplete, or inaccurate data is likely to lead to mistaken conclusions and decisions. The old adage, that 90% of a data analysis should be spent on cleaning the data applies even more so with very large data sets.

3. **Algorithm development.** This includes developing the algorithm and testing and validating it. Key questions should be asked at this stage. They include: is the objective function optimised by the algorithm really the one we want to optimise? Are there bugs in the software which manifest themselves only in unusual conditions? If software packages are used, are we confident that their default settings are doing what we want?

4. **Deployment, monitoring, and maintenance.** AI assurance does not stop once the data has been captured, the system built, and the algorithm developed. Its performance needs to be evaluated, monitored, and audited. It is important to regularly check whether it is performing well enough. We need to know how it performs when circumstances change – after all, the one thing we know about human society is that change is constant. How does the system respond to anomalous conditions or data it has not previously encountered. What fall-back plans are in place should the system go down? It might be appropriate to update the system to broaden its capabilities when inadequacies are detected – but care has to be taken that unexpected interactions do not lead to overall degradation of performance.

5. **Legal requirements.** It should go without saying that the system must adhere to data governance legislation and regulation. It is imperative also to consider whether there is

adequate human oversight, which is sometimes a legal requirement. This stage also includes security considerations and resilience to attack, fraud, and intrusion attempts.

6. **Ethical considerations**. This will include issues of discrimination, transparency, and privacy preservation, all of which are also often legal requirements. It will also include considerations of the wider impact on society – we must learn from the adverse effects of social media.

AI technology promises to significantly enhance the human condition. However, as with any advanced technology, it comes with risks. The risk mitigation strategy described above can give us confidence in the future.

More explicitly, our position on the further development of AI:-

1. **Responsibility and accountability**. Organizations that develop and deploy AI must be fully responsible and accountable for the consequences of those systems. We recommend the creation of a role such as AI Officer who will take day-to-day responsibility for monitoring AI developments and risks within the organization. This is in analogy to the Data Protection Officer role required as part of European Union General Data Protection Regulation (GDPR). Responsibility for deployment of trusted AI should still be at all levels of the organization, in particular, senior management, but having an AI Officer will allow an organization to be more able to manage AI risks. We encourage governance and support regulation, especially for high stakes applications of AI such as in medical and social contexts.

2. **Code of Practice.** It is necessary to ensure that AI systems work in a way that is of benefit to their human users and perform the tasks that they were designed to perform, without negative side-effects. To this end there is a requirement for the creation of practitioner-centric *codes of practice* that AI developers need to follow to ensure that AI systems can be trusted. *Validate AI* will be involved in this process and will be in discussion with other agencies and bodies that are also working towards this goal.

3. **Convening.** It is important to draw together the various parties who are interested in the development of AI, such as those businesses and practitioners who wish to innovate and take advantage of the rapid development of AI technology, those who are more cautious about the risks of AI, and those communities that perceive a risk to them with AI development such as loss of employment. We promote dialogue amongst these different groups. *Validate AI* is an apolitical and impartial trusted third party.

4. **Independent audit.** AI systems may be deployed without independent checks and this is a major source of risk. We maintain that high impact AI systems should only be deployed following independent and rigorous assessment and auditing, as is the case with many other activities such as medical services, airlines, agriculture and food safety. To this end, a new profession for AI assurance is required and supported by *Validate AI*.

5. **Monitoring.** AI systems are prone to unexpected behaviour; therefore it is important that even once deployed, AI systems are carefully monitored and resources are set aside to do this. Contingency plans need to be in place to deal with scenarios when AI could fail in the future, to reduce the negative impact of AI failures. This requires an understanding of the lifecycle of an AI

system and its reliance on historic data. In particular, this is true also of reliance on generative AI systems that may become out-of-date quickly since they are trained on old data.

6. **Education.** The more we can educate businesses, IT and AI developers, the government and the general public about AI, how it works, what the risks are, and what to expect, the more able we will all be to anticipate and mitigate the risks of AI, and the more able businesses and the general public will be in assessing and using AI systems. Education should be practitioner-centric, providing AI developers with tools and knowledge they can use directly in their development roles. For more general education about AI, this should be tailored to different roles in business, public service, government or the public. *Validate AI* will continue to be involved in this education process, arranging conferences, workshops and training programmes, and commissioning specialized white papers.

Annex: A Note on Testing AI Systems

The dramatic advances in machine learning and AI systems that we have witnessed in recent years are primarily a consequence of a switch from deductive to inductive systems. The former are based on the logical application of well-defined rules to clear premises. The latter are based on identifying structures and patterns in data. The development of powerful computers, coupled with the availability of massive data sets, has given a tremendous impetus to systems of the second kind, as we see in deep learning and chatbots and generative AI more widely.

Key concepts are *trustworthiness* and *explainability*. A system is trustworthy if there is evidence that it reliably produces accurate results – that its conclusions can be trusted. A system which can describe the steps leading to its conclusion, that is a system which is *transparent*, demonstrates explainability and contributes to its trustworthiness.

Formal verification is difficult in the case of inductive AI but tools exist to explore which aspects of the data are especially important in a decision, to measure success rate, to cast uncertain cases back to a human, and which permit questions to be modified or adjusted in the light of earlier responses, so refining the question and achieving a more accurate answer.

Clearly much of this is highly application-context dependent and potential users need to ask themselves what would count as an adequate system, what extent and type of interpretability is required, and how much training is needed for people to use the system effectively.

Contact us at:

Contact us at: contact@validateai.org
Website: www.validateai.org