The Journey to Trustworthy AI – Part 1:
Pursuit of Pragmatic Frameworks

Mohamad M. Nasr-Azadani & Jean-Luc Chatelain

Date: March 8, 2024

Abstract

This paper, the first installment in a series on **Trustworthy Artificial Intelligence (TAI)**, reviews various definitions of TAI– and its *extended family*. Considering the **principles** respected in any society, TAI is often characterized through a range of attributes or subjective concepts, some of which have led to confusion in regulatory and engineering contexts. We argue against the use of terms such as *Responsible* or *Ethical* AI to substitute TAI. And to help clarify any confusion, we suggest leaving them all behind. Given the subjectivity and complexity inherent in TAI, developing a universal framework is deemed infeasible. Instead, we advocate any approach centered on addressing key attributes and properties such as *fairness*, *bias*, *risk*, *security*, *explainability*, and *reliability*. We examine the ongoing regulatory landscape, with a focus on initiatives in the European Union, China, and the USA. We recognize that differences in AI regulations based on geopolitical and geographical reasons pose an additional challenge for multinational companies. We identify *risk* as a core principle in AI regulation and TAI. For example, as outlined in the EU-AI Act, organizations must gauge the *risk level* of their AI products and act accordingly (or risk hefty fines).

We compare common modalities of TAI implementation and how multiple cross-functional teams are engaged in the end-to-end process of TAI for any organization. Thus, a brute force approach for enacting TAI renders efficiency and agility, moot. To address this, we introduce our framework '*Set–Formalize–Measure–Act*' (SFMA). Our solution highlights the importance of transforming TAI-aware metrics, drivers of TAI, stakeholders, and business/legal requirements into actual benchmarks or tests. Finally, overregulation driven by panic of powerful AI models can, in fact, harm TAI too. Based on GitHub user-activity data, in 2023, AI open-source projects *rose to top projects* by contributor account. Enabling innovation in AI and TAI hinges on independent contributions of the open-source community.

Contents

Summary Points

Key Takeaways

- Trustworthy AI (TAI) is an evolving concept.
- There is no '*one-size-fits-all*' solution for TAI.
- AI will have impacted human civilization at scales not fully understood yet.
- Meanwhile, there is no need to either Panic or underestimate the impact of AI, globally.
- The viable path towards TAI would involve collaboration amongst communities, regulators, private sector, open-source communities, academia, and legal scholars (to name a few).
- Open-source Software movement has been fueling innovation for decades. Let's nurture (and not restrict) it to lead actualization of TAI tooling.
- Experts across various disciplines can play a key role in "translating" principles of TAI into "attributes" or "properties" such as: safety, reliability, fairness, explainability…
- There is no single universal framework that can deliver TAI in an organization. Instead, we suggest communities focus on appropriate definition and measurement of relevant metrics for any TAI attribute.
- Several regulatory bodies such as the European Union have approached TAI from a risk management perspective.
- Clear understanding of uncertainties in AI models life cycle should be mapped to appropriate risk management frameworks such as Rumsfeld Matrix. This can enable decision-makers with tools to face uncertainty.
- Terms such as 'fairness', 'bias', 'accountability', 'ethical' are loaded concepts with deep roots in any community or country's culture, history, societal values, and governments.
- Association of these terms as 'principles' of TAI is ultimately context dependent and requires careful 'infusion' into regulatory and engineering systems.
- Mathematically speaking, it has been demonstrated that it is impossible to honor every manifestation or aspect of AI-fairness simultaneously.

Disclaimer

☞ While discussions in this write-up aims for longevity, AI regulations and legislation are still evolving in many countries. Therefore, some discussions may require updates as new regulations emerge. [Last updated: March 8, 2024]

✂ In this work, we do not consider the following AI-systems:

- ✗ **AI-controlled and fully-autonomous robotic systems:** For a recent survey, cf. Ingrand and Ghallab (2017); Kunze et al. (2018).
- ✗ **In vivo AI-powered synthetic biology and biotechnology:** For example, *Xenobots* (cf. Blackiston et al. (2021) and Kriegman et al. (2021)).
- ✗ **Self-evolving and self-replicating AI:** For example, cf. News (2023).
- ✗ **Quantum machine learning:** For a recent survey, cf. Zhang and Ni (2020).

1 Context

Across the globe, many governments and legislative bodies are actively working to regulate the development and use of 'Artificial Intelligence' (AI), cf. Reuters (2023). For instance, President Biden's recent executive order on 'safe, secure, and trustworthy AI' issued in October 2023[1] was quickly followed by a similar announcement from the 'European Union' (EU) in which the EU members unanimously reached a *political agreement* to regulate AI[2] (European Parliament Press (2023)).

The primary impetus behind the ongoing regulation of AI is the wide-ranging impact it will have on every facet of human life. In essence, AI in conjunction with existing/emerging technologies such as 'Internet of Things' (IoT), 5G/6G[3], and 'Digital transformation' (DX) is poised to impact the so-called 'Fourth Industrial Revolution' (4IR), cf. Philbeck and Davis (2018); French et al. (2021).

In general, *risk* and *uncertainty* are considered *intrinsic properties* in many autonomous systems. AI-powered systems are not exempt from this classification either. Hence, the term 'Trustworthy Artificial Intelligence' (TAI) has been coined, representing multi-disciplinary research areas tackling the '*distrust*' in AI systems. With the remarkable performance of recent AI products, such as ChatGPT, regulatory bodies have accelerated their efforts in passing legislation. While we recognize substantiated concerns raised by public and prominent research scholars[4], we warn against the over-regulation of AI. In several cases, open-source

[1] Executive Order (EO)-14110 was first released by the White House on October 30th, 2023. Full draft and text can be found here: Biden (2023).

[2] Commonly known as the '**EU-AI Act**', this legislation is expected to go into effect in 2025 or 2026 and has been hailed by many as the first comprehensive legislation for TAI.

[3] 6G telecommunication network paradigms are still in the research stage, cf. Jiang et al. (2021). 6G's overall mission is to build the communication platform which a hybrid world consisting of physical and digital realities, e.g. 'Augmented Reality' (AR), can function, seamlessly. Commercial 6G is expected to arrive in late 2020s or early 2030s, cf. Telefonaktiebolaget LM Ericsson.

[4] Recently, in response to remarkable *human-like* capabilities demonstrated by a new family of AI models called 'Generative Pre-trained Transformer' (GPT) and 'Large Language Model' (LLM), public opinion along with that of prominent AI academics such as Professor Geoffrey Hinton, has raised serious concerns about the potential existential threat posed by AI models, e.g. Barrat (2023). While we do not discount the possibility of *doomsday events* triggered by 'AI-*gone-wrong*', addressing circumstances that can lead to catastrophes of such magnitude is beyond the scope of this article. For more on this topic, we refer reader to recent surveys, cf. Galanos (2019); Carlsmith (2022); Bucknall and Dori-Hacohen (2022); Federspiel et al. (2023).

community is being heavily targeted which could hinder– the much needed– innovation from such communities to deliver TAI. We will discuss this in § 9.4.

Considering efforts to regulate AI, we argue that TAI– along with the disciplines surrounding it–has played a unique role in the path ahead: It has motivated cross-functional collaboration amongst experts and stakeholders, and regulatory entities to:

- ☞ Understand cutting edge AI technologies,
- ☞ Assess the near- and long-term impact of AI on society and the economy
- ☞ Propose new policies, standards, and frameworks,
- ☞ Solicit and incorporate feedback from the public domain in TAI policies,
- ☞ Enact new or revise existing laws, standards, and guidelines.

In this work, we hope to provide our point of view on '*journey*' towards the realization of TAI. In doing so, in part 1, we demonstrate how numerous 'principles' of TAI[5] could be aggregated and be "*transformed*" into tangible 'frameworks' enabling TAI within any organization.

We provide a summary of characterizations as well as *taxonomy* used by multi-disciplinary scholars addressing TAI and its derivatives, e.g. 'eXplainable Artificial Intelligence' (XAI) or AI fairness. In § 8.4 we introduce our proposed solution called '***Set, Formalize, Measure, and Act***', a simple yet powerful framework towards TAI for enterprise.

In part 2, we provide an overview of recent advancements in statistical and data-driven techniques for quantifying critical metrics representing every dimension of TAI. We will compare different modalities of implementing TAI, prioritizing 'Trustworthy-By-Design' frameworks.

2 Trustworthy AI: Too Many Definitions or Lack Thereof?

We argue that there is not concrete definition for the terminology 'trustworthy artificial intelligence' insofar as it has been characterized by the desired attributes in a particular discipline such as engineering, education, economy and markets, and public policy, cf. Stix (2022).

Table 1: Attributes extracted from principles defining 'Trustworthy AI' that have been announced by various entities. For a complete list of principles for each, see appendix A.2.

Entity Name	Framework or Theme	Safe Secure	Privacy-enhanced	Explainable Interpretable	Transparent Accountable	Human Oversight	Robust Resilient	Reliable Valid	Prioritizing Humans	Fair	Literacy Awareness
NIST	Risk Management	✓	✓		✓		✓	✓			
UNESCO	Human Rights	✓	✓	✓	✓	✓	✓		✓	✓	✓
IEEE	Trustworthy-by-Design				✓	✓			✓		✓
OECD	Democracy & Market Economy	✓		✓	✓		✓		✓	✓	

As such, any entity that aims to define TAI should consider factors such as applications (or services), business goals, legal context, and parties involved, amongst other important elements (for a recent review on TAI definitions and taxonomy, we refer the reader to Thiebes et al. (2021); Jacovi et al. (2021)).

[5] So far, several domestic and international organizations have released lists of '***TAI mission.***'. While there are common items their compiled lists, we emphasize that every organization prioritizes a certain aspect of human life and society that is aligned with its mission when publishing principles of TAI. For example, IEEE is focused on building robust standards for engineering applications. Alternatively, UNESCO is focused on human rights and education. We will discuss these later in the text.

In the remainder, we recap the principles for TAI recommended by various governmental and other international entities, namely the 'National Institute of Standards and Technology' (NIST), the 'United Nation's Educational, Scientific and Cultural Organization' (UNESCO), the 'Institute of Electrical and Electronics Engineers' (IEEE), and the 'Organization for Economic Co-operation and Development' (OECD).

We have selected this diverse list of entities to demonstrate the commonalities in TAI principles despite their varying missions. In table 1, we provide an aggregated view to demonstrate how independent international or domestic units focus on the various *features* in an AI system to be considered a TAI. For more details on the principles for each entity, see appendix A.2.

2.1 Trustworthy AI: *Attribute* or *Property*?

The process of building TAI very quickly turned into an amalgamation of an ever-growing attributes expected from an AI system. For example, '*Fairness in AI*', '*AI Safety*', '*Secure AI*', '*Transparent AI*', '*Explainable AI*', '*Interpretable AI*', '*Black-box AI*', '*Responsible AI*', '*Robust AI*', '*Resilient AI*', '*Ethical AI*', '*Reliable AI*', '*Privacy-enhanced AI*', '*Accountable AI*', and '*Federated AI*' are common examples for an "extended" AI family. In other words, and out of necessity, we have been "cooking" this topic in a *magic pot* with "chefs" from various disciplines.

We must keep in mind that the majority of the aforementioned terms characterizing TAI do not possess a universally accepted definition. A few terms are used interchangeably. For example, consider 'interpretability' and 'explainability' that are used synonymously. From an engineering perspective, 'explainable AI' and 'interpretable AI' point to two distinct technical concepts. For instance, ***outputs*** returned by a 'Deep Neural Network' (DNN) model can be ***explained*** using algorithms such as LIME (Ribeiro et al. (2016)) despite DNNs categorized as not interpretable[6]. In contrast, it is widely accepted that term 'interpretability' should be classified as an intrinsic property when selecting a family of AI model. For example, deploying a 'Decision Tree Classifier' as an AI product provides 'interpretability' almost at not additional compute cost. This is due to its inherent '***If-Then-Else***' topology when computing an outcome. To summarize, every feature (attribute) utilized to characterize TAI is either:

I. **An Intrinsic Property:** An inherent attribute or a characteristic of an *object* which does not depend on its external environment, relationship, or conditions. Hardness and mass are intrinsic properties of a diamond. We argue that classifying properties of a TAI system is necessary and can simplify the frameworks, legal ramifications, and implementation techniques. An example of an intrinsic property in an AI product is its degree of "black-boxness". In this context, 'Black-box' AI– a term predominantly used by AI engineers– is an *intrinsic property* of an AI model. It indicates a category of AI model where the underlying mathematical reasoning is non-linear and complex to be readily understood by humans.

II. **An Extrinsic Property:** An extrinsic property on an object or substance depends on the external factors and relationships with other external objects. For instance, temperature of an object depends on the surrounding environment. In an AI system, such properties can be assumed an 'add-on' to an existing AI model. For instance,

[6] In literature, DNNs are categorized as 'Black-box' AI models.

an AI team can make an existing computer vision model 'secure' by adding additional layers to it could be deployed in a high-risk use-case such as self-driving cars. In other words, a company could initially train a sophisticated and reliable computer vision model without implementing 'security', and subsequently apply methods to add this (extrinsic) property in an on-demand manner.

Next, one may ask why these categories matter? Without going into details, agreeing on such classification clearly early on could help any organization with implementing and maintaining TAI in its product life-cycle[7]. Consider the common yet important decisions impacting AI product lifecycle.

- ☞ **Metric Selection:** Map requirements to metrics. For example, there are numerous ways to evaluate '*fairness*' of a loan approval AI model. Selection and compute the 'fairness-score' may not be trivial.
- ☞ **Enterprise Risk Management (ERM):** Since any data-driven product inherently is not '*bullet-proof*', risk-assessment and management frameworks used currently by an enterprise can impact the 'Uncertainty Quantification' (UQ) techniques which may be directly tied to ERM.
- ☞ **Resource Allocation:** Allocate and plan on resources such as human 'Subject Matter Expert' (SME), continuous monitoring and improvement platforms for AI systems running in production.
- ☞ **Legal Compliance:** Understanding the risks involved in violating legal obligations is the first step to plan and absorb the inherent risk associated with any AI-product.

2.2 Challenges-turned-into Myths Surrounding Trustworthy AI

It is fascinating to watch how the topic of Trustworthy AI– and its variants– has been debated by scholars and policy makers across a wide number of domains. Several scholars argue that assigning terms such as 'trustworthy' or 'responsible' to AI (in the context of legislation) may confuse various sectors. If not properly differentiated, it ultimately undermines proper implementation and enforcement of TAI, cf. Freiman (2023); Laux et al. (2024).

2.2.1 Example Myths about Trustworthy AI

To bring clarity surrounding TAI and its definitions, it is important to, first, recognize questions or assumptions that eventually rendered technical challenges as *myths*. Here are a few examples:

- ✗ **Myth:** Products using AI are autonomous; therefore, their "*decisions*" cannot be comprehended or defended.
- ✗ **Myth:** We (humans) are not capable of rationalizing the decisions made by black-box AI models.
- ✗ **Myth:** We cannot "*control*" the decisions of an AI system.
- ✗ **Myth:** The only *reason* for an AI model to act ***unethically*** is due to its training performed by a human (or a human-supervised system).

[7] Unlike the mature software development life-cycle, as of today, there seems to be no universally accepted framework for the life-cycle of AI products. This is partially due to their dependance on organizational structure, business processes, and the mode of AI integration within the enterprise.

- ✗ **Myth:** Any AI model that is trained on real-world data– echoing human history, values, and the evolution of society– cannot have its harmful biases mitigated.
- ✗ **Myth:** Any decision made solely based on *human intuition* always outperforms than that of an AI system (or *vice versa*).
- ✗ **Myth:** With the emergence of larger and more powerful AI models, e.g. ChatGPT, humans are to be completely removed from the decision-making process.[8]

While having healthy debates around these topics or myths is always welcome, it should not promote valid concerns into paralyzing or panic shutting down progress in AI.

2.3 Trusting AI Systems: A Complicated Relationship with Humans

"*In republics, the people give their favor, never their trust.*" (Antoine Rivarol (1753–1801); A French writer)

One might simply ask: 'What is trust?'.
To make matters more complicated, there is no unified definition for '***trust***' across different disciplines. Psychologists consider trust a **cognitive attribute of the human mind**[9], sociologists associate trust with **human relationships**[10], and economists argue that trust[11] can, in fact, be '**calculated**' (Granovetter (2018)). For a comprehensive list of definitions for *trust* across various disciplines, we refer reader to (Cho et al., 2015) and references therein.

The presence and influence of decisions made by automatic algorithmic systems is undeniable. Recently, terms such as '*algocracy*' (algorithmic government) have been used to describe potential 'futuristic' governments. Such ideas are not far-fetched. For example, a software named COMPAS is used in justice systems in the USA to help judges assess the likelihood that a defendant becomes a recidivist (we discuss this in § 6.6). Additionally, it is estimated that the majority of trading performed on Wall Street is carried out by autonomous algorithms and trading bots, cf. Patterson (2013); Menkveld (2016); Isidore (2018).

2.3.1 How do we (Humans) Trust the *Unknown*: It is always a Process.

As history has shown us, when faced with new technologies such as *electricity*, *the Microwave oven*, or AI, many people typically respond with *justified* skepticism, resistance[12], fear, and

[8] Currently, it is widely accepted that the human brain outperforms the '*best* ' 'Artificial General Intelligence' (AGI) system. Qualities such as '*out-of-the-box* ' thinking, and '*causal reasoning* ' (Bishop (2021)) are considered example 'super-powers' of human brain. There is a strong consensus in the scientific community that by only increasing the ***size*** and enhancing the ***capacity*** of AI models, we cannot produce AGIs capable of outsmarting humans in every capability, cf. LeCun (2023); Fjelland (2020).

[9] Rotter (1980) defines trust as: '*Cognitive learning process obtained from social experiences based on the consequences of trusting behaviors'.*

[10] As in sociology, trust is defined as (Gambetta et al. (2000)): '*Subjective probability that another party will perform an action that will not hurt my interest under uncertainty and ignorance'.*

[11] In James Jr (2002) and in the context of economic systems, trust is defined as: '*Expectation upon a risky action under uncertainty and ignorance based on the calculated incentives for the action*'.

[12] An example is the *Printing Press* introduced in the 15th century, which faced resistance from Catholic Church as well as monarchies in Europe. Such entities relied on censorship, manipulated licensing systems, and enforced heavy penalties for 'unapproved printing' to limit the impact of the printing press on educating people. With education becoming more accessible to a wider audience, the control of religious rulers, governments, and monarchs over the people was jeopardized, cf. Pardue (2012); Robertson (2015).

sometimes, complete backlash against innovations like 'Google Glass' (Kudina and Verbeek (2019)).

To overcome such resistance, it is important to harness the power of '*trust*'. The successful interplay of 'Realizing Trust' and 'Human Societies' commonly undergoes several steps (cf. Frischmann and Selinger (2018); Lankton et al. (2015)) listed below:

a) **Establishing Trust:** Properly and transparently 'introduce' new technology to the community. In addition, 'educate' users on how to interact and utilize it.

b) **Building Trust:** Allow users interact with the new technology in a safe and guided manner. When many users *consistently* have positive experience in their engagement with the new system and notice that the outcomes align with their '*ethical*' norms, it can be assumed that *trust is built*.

c) **Maintaining Trust:** Requires ongoing effort to ensure continuous improvement, demonstrating willingness and adaptability to evolving challenges, and open and honest communication channels with their users.

d) **Rebuilding Trust (If needed):** As no system is perfect, when a new system fails, restore and rebuilding trust requires steps to remediate the problem, remove any culprit(s), and be transparent with its users upon completion of conducted 'Root Cause Analysis' (RCA).

e) **Sustaining Trust:** Requires steps to encourage the involvement of communities in the long-term engagements and fostering the technology at hand by providing feedback channels and a focus on long term value.

Without delving into specifics, we note that the process of building trust between 'an individual person' (as opposed to a group or a community) and a new technology can differ from steps discussed above. Psychological and biological variations could significantly influence the outcome.

2.3.2 UK Home Office's Biased Algorithm: An Example of Failure in Building Trust

The UK Home Office faced criticism for its use of an AI algorithmic system in processing visa applications, which came to light in 2018, cf. Gualdi and Cordella (2021). Before it was publicly labeled as a biased and 'racist algorithm' BBC News (2020), this AI engine had been built to "*streamline*" the heavily backlogged visa application process. Towards this, given a visa applicant, this AI product "categorized" applications into various risk levels and identify "high-risk cases" for further scrutiny.

Let's recap the challenges and how actions (or lack thereof) damaged 'Trust' between immigrant communities and the UK Home office:

(a) **Familiarity and Consistency:** The introduction of the algorithm disrupted the familiarity for visa applicants, as 'black-box' and automated system suddenly played a significant role in the decision-making process.

(b) **Transparency:** The new algorithm lacked *transparency* in its decision-making process.

> "*Potentially life-changing decisions are partly made by a computer program, that nobody on the outside was permitted to see or to test*", Cori Crider, Foxglove (Katie Collins–CNET (2020)).

Visa applicants were not informed about the criteria and factors used by this algorithm which determined their "*risk level* ", leading to concerns about accountability of the system as a whole.

(c) **Perceived Competence:** Concerns about the origin of the new algorithm (and its training dataset), accuracy, and fairness risk scores raised questions about the competence of the UK Home Office in overseeing and implementing its new Visa processing system using AI.

> "*... Researchers from Foxglove and the JCWI believed it was built in house by the {UK} government rather than brought in from a private company. They allege that the government is being purposefully opaque about the algorithm because it discriminates based on the nationality of the applicant, and that it doesn't want to release a list of the countries it considers high risk into the public domain.*" (Katie Collins–CNET (2020))

(d) **User Control:** Visa applicants (or independent legal entities) had limited "control" (if any) over the decision-making process. The lack of transparency did not allow them to address issues or to petition a decision made by the UK Home Office in a meaningful manner– in an event of a rejection outcome.

(e) **Long-Term Relationship Building:** Trust issues stemming from the opacity (lack thereof) of the algorithmic decision-making process potentially harmed the long-term relationship between the government and visa applicants.

> "*We also discovered that the algorithm suffered from 'feedback loop' problems known to plague many such automated systems - where past bias and discrimination, fed into a computer program, reinforce future bias and discrimination. Researchers documented this issue with predictive policing systems in the US, and we realised the same problem had crept in here.*" (Foxglove (2020))

Given the circumstances, rebuilding trust requires addressing concerns from all parties involved, increasing transparency in existing or future models, and providing avenues public-facing auditing mechanisms.

(f) **Community Involvement:** This automated system–which was in place for five years– incorrectly rejected numerous visa applications to the UK based solely on the *applicant's country of origin*. This could have been remediated earlier if immigration advocacy groups, independent technical firms, legal councils, and applicants, were all included in discussions and oversight about the use of automated decision-making tools.

3 Complexities and Challenges

Key Takeaways

- ❒ Let's avoid a 'wild-goose chase': AI is not the "*responsible*" agent in the room: Its users and companies are.

In this section, we aim to provide our insight on why there is no 'one-size-fits-all' solution for TAI.

3.1 "*Responsible AI*": A Confusing Term that should be Left Behind

It is safe to assume that by now, AI and related disciplines such as 'Machine Learning' (ML) or Data Science, are independent scientific paradigms, akin to mathematics or statistics. Just as no one expects to comprehend phrases such as 'Responsible Mathematics', we argue that the term 'Responsible AI' is meaningless. Since AI has already become an essential tool for aiding product teams, it is actors how decide how to utilize it in their business. Simply put, 'Irresponsible actors/engineers/managers' do exist, not so much 'Responsible AI' and its evil twin, 'Irresponsible AI'.

3.1.1 Mathematics cannot be Held "Responsible", nor should AI

Mathematics, inherently, cannot be held accountable; rather, the responsibility lies with the individual or entity utilizing mathematics. In a similar vein, same principles apply to other scientific disciplines including AI. Without digging into current legal and philosophical debates surrounding agency as well as accountability surrounding any autonomous system, in scenarios where an AI-product '*is*' in charge of making decisions autonomously and independently, entity who passed on such responsibility to this product would be held liable.

3.1.2 Examples: Achieving Clarity by not Expecting a Product to be "Responsible"

To drive our point home, let's imagine we encounter news headlines such as the following list:

- ✗ MagicKar, a car manufacturing company, is making a '*Responsible Self-driving Car*' as their next model.
- ✗ President of University of MarsY forms a committee to develop a framework for '*Responsible Computer Science*'.
- ✗ An online search engine company, called Tix-Tax-Tox, announces the release of its new '*Responsible Search Engine*'.

Statements above while *grammatically* correct, are not semantically comprehensible– to say the least. Any organization tasked to build a 'responsible product X' will have follow-up questions such as '*a) What is considered a responsible car? or b) Is this a legal or ethical mandate?* '. In response to such clarifying questions, a person has to only use context-aware and relevant terms to describe 'being responsible or 'acting responsible':

- ✓ ..., MagicKar, is making a '~~*Responsible*~~ ***Safe*** *Self-driving Car*' as their next model.
- ✓ ..., a committee to develop a framework for '~~*Responsible*~~ ***Transparent & Resilient*** *use of Computer Science*'.
- ✓ ..., company announces the release of its new '~~*Responsible*~~ ***Unbiased*** *Search Engine*'.

3.1.3 An Undesirable Outcome for AI Industry: "*Responsibility-as-a-Service*"

The title says it all... Considering the complexities of TAI and soon-to-be-enacted AI regulations, this scenario may occur seamlessly– if not already. Attaching 'Responsible' as a characteristic of an AI system could marginalize the significant effort, technical debt, legal considerations, and human expertise required. Driven by a highly competitive market in AI, we observe signs of such shift in building large scale AI-enabled products: In essence, a company first trains an AI model only focusing on its **performance** and **accuracy**. Once model

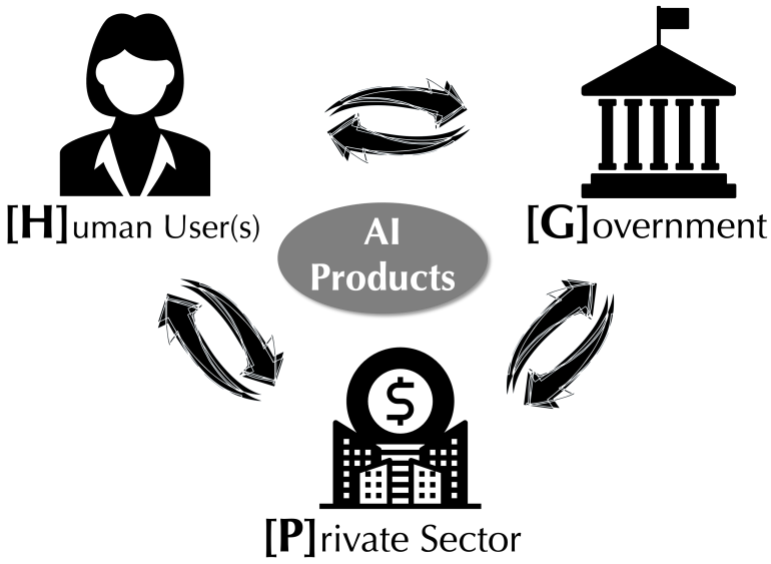

Figure 1: Main parties involved in assessing 'Trustworthy AI' in a product or service; **H**uman end user (or community); **G**overnment; and the **P**rivate sector. Note that for every two entities, any acceptable TAI framework should be equipped to address the any professional (two-way) interactions.

is trained, this same company attempt to find out what and how it can **make** it '*responsible*' without deteriorating the AI model's accuracy– as long as the upgraded AI model somehow can remain within the legal bounds. If bounds are relaxed or became more stringent, this company would only expand or shrink their team or resources allocated for "RaaS-AI", accordingly.

We hope that we have convinced you that the term 'Responsible AI' is not a suitable ambassador for TAI, hitherto. This is further pronounced in legislation and regulatory contexts. Practically speaking, proper integration and usage of AI models in any product, application, or services approved by governing bodies, ought to be carried out following a multi-tiered legislative or regulatory enforcement. Many countries have recently started experimenting multi-tiered regulation of AI. Some even have set *risk* as the core element in their TAI regulatory frameworks. We discuss this further in § 4.

3.2 Trust and the Parties Involved

Figure 1 shows three distinctive entity type that can interact in a business or professional context. In essence, either one- or two-way interactions[13], need be considered when a target TAI framework is to be developed.
Below we categorize varieties of 'two-way interactions' (see fig. 1) that can occur in any professional or social context:

- H ⇆ G: Human interactions with Government (and *vice versa*). Example: Use of AI by judiciary system and the rights of citizens.
- H ⇆ P: Human interactions with Private entities (and *vice versa*). Example: Use of AI by a bank to approve/reject a citizen's loan application.

[13] Note that there are other possible categories, e.g. self-self and three-way interactions. For the sake of simplicity, we do not discuss them here.

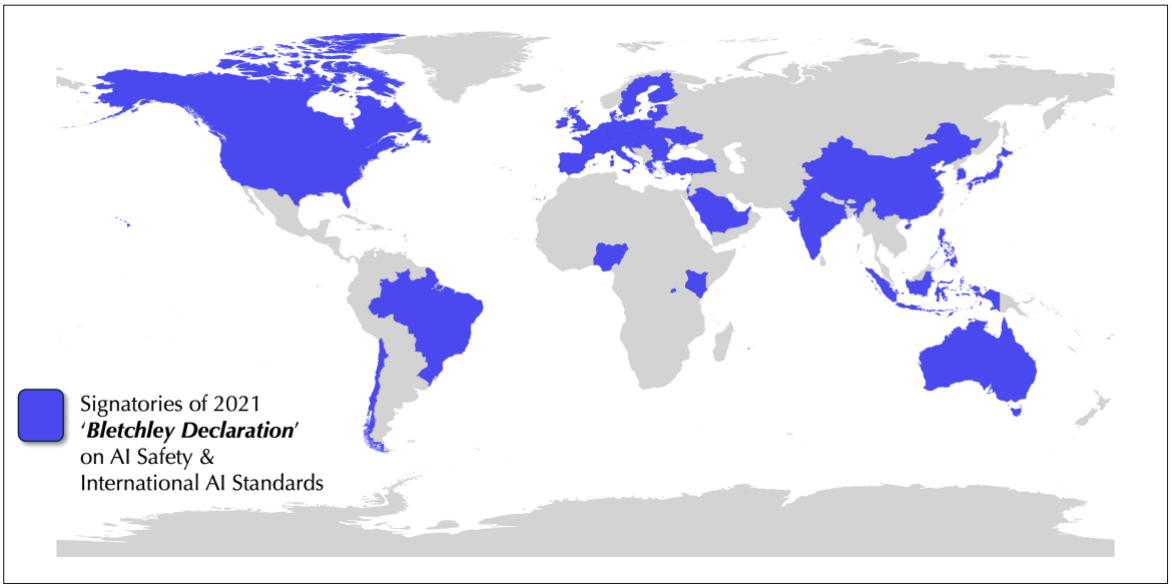

Figure 2: The first international summit on AI Safety held in November 2023 in Bletchley, UK. Twenty-eight countries signed 'Bletchley Declaration'. List of countries retrieved from (Toney and Probasco, 2023).

- $G \leftrightarrows P$: Government interactions with private entities (and *vice versa*). Example: Use of AI by 'Federal Trade Commission' (FTC) to investigate reports of illegal activities carried out by a specific bank.
- $G \leftrightarrows G'$: One government entity interacting with another government entity (and *vice versa*). Example: The Supreme Court of USA investigating data-backed claims regarding 'gerrymandering' in a particular state.
- $H \leftrightarrows H'$: Human interacting with another human. Example: A citizen using AI to publish fake images of a former colleague.
- $P \leftrightarrows P'$: Private entity interacting with another private entity. Example: An internet search engine giant throttling internet speed only for iPhone (as opposed to Android) users.

Attributes associated with TAI are directly or indirectly be impacted by the family of inter- action and entities involved. For example, explainability– a pillar in any TAI framework– requirements are different for a government's legal investigation *vs* a social media user requesting explainability on how her activity data was used to see particular advertisements.

3.3 Geographical and Geopolitical Considerations

The first international conference called 'AI Safety Summit' was held in the United Kingdom in November 2023. This event concluded with 28 countries signing an agreement known as the '**Bletchley Declaration**' (see fig. 2). First of its kind, Bletchley Declaration focuses on the challenges and risks of AI and, therefore, seeks cooperation among international communities and countries to establish cooperating channels to mitigate risks posed by AI (Government of the United Kingdom (2023)). While Bletchley Declaration is a good example of international cooperation to regulate AI, geopolitical dynamics play an important role in making or breaking such efforts.

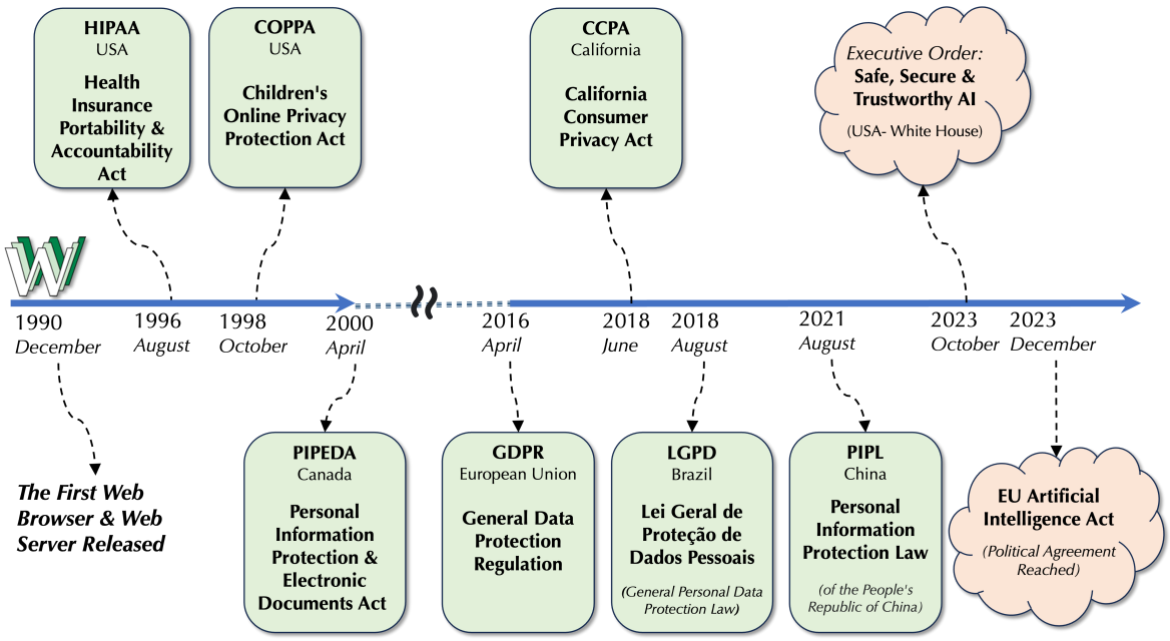

Figure 3: Timeline of data privacy laws passed by example legislative entities. While many countries have not yet passed or enacted their digital data privacy laws, public opinion is now pressring them to pass laws to regulate AI. For reference, in 1990, the first web sever and web browser were created by *Sir Tim Berners-Lee*.

Consider leading global powers such as the United States, China, and the European Union. The United States, with its tech giants and well-established innovation ecosystems, sets critical trends in the development of AI and TAI centered on markets, while China's state-driven approach where it prioritizes a centralized authority to regulating AI in areas such as *content generation* or *recommendation systems*. The EU, however, following its existing strict data privacy and ethical standards such as GDPR, is now taking a strict approach to regulate AI through with 'risk' at its core (we discuss this in § 4).
In summary, varying approaches to AI governance at regional and international scales are shaped by factors such as political and technological leadership, data sovereignty laws, cyber- security threats, cultural as well as ethical perspectives on AI use (see fig. 4.

3.4 AI Regulation Modes: Bottom-up *vs* Top-down Development

As far as the modality of AI regulation is concerned, governments and international organizations have been experimenting with different implementation approaches. Common frame- works on regulation and governance are as follows:

- ☞ **Top-down Regulation:** Rules set by higher authorities or central government, trickle down to ensure compliance. It is widely used across sectors like finance, healthcare, and telecommunications to maintain order and public safety. Critics of this approach argue it stifles innovation and adaptability (Homsy et al. (2019)).

- ☞ **Bottom-up Regulation:** In contrast to 'top-down regulation', this approach starts from local communities and governance. It heavily relies on self-regulation as well as community governance driven by grassroots organizations and independent entities.
  The 'flow' of regulation is, therefore, upwards with higher authorities adopting the collectively verified policies (Capano et al. (2012)).

- ☞ **Multi-level Regulation & Governance:** Multilevel governance recognizes that policies and rules must be flexible enough to be adopted at various levels, e.g. local, regional, national, and international levels. It involves coordination and cooperation among these different levels of government, as well as with non-state actors, to address issues that may cross traditional jurisdictional boundaries, e.g. mitigating risks of climate change (Tortola (2017)).

- ☞ **Other Forms:** For instance, *Market-based Regulation*, *Self-regulation*, *Horizontal Regulation & Governance*, *Network-based Governance*, *Democratic Governance*, and *Hybrid Modes* are a few examples. For a review, we refer reader to Levi-Faur (2012)).

One major concern raised is how innovation in AI innovation may be impacted by the choice above? There is a clear **trade-off** between the level of regulation[14] and innovation (Chan et al. (2022)). In other words, a top-down approach may seem a natural way to start regulating AI where a central governmental entity '*defines and controls*' enforcement. We have observed signs of such viewpoints in EU-AI Act requesting permission to use AI in certain 'high-risk' domains (see § 4). Alternatively, a bottom-up governance heavily relies on private sector to "*self-regulate*" and follow 'best practices' in AI products and ecosystem[15].

3.4.1 A Few Open Questions

For policy makers or communities aiming to be involved in regulation of AI and developing TAI frameworks, answers on the following ought to be considered:

- ☞ Should TAI and its legislation be based on a top-down, bottom-up, or market first?
- ☞ Can we prioritize a *bottom-up* strategy, involve STEM academics, social sciences, and legal scholars to lead debates and building the legal framework?
- ☞ Alternatively, prioritize government's role and authority in passing TAI regulations.
- ☞ Should any TAI framework be accepted and adopted within international communities, first?
- ☞ Should regulation of AI be approached through only a lens of '*risk*', '*security*', '*national security*', '*social/criminal justice*', '*commerce*', '*human rights*', '*social prosperity*', '*existential threat to humanity*', etc. ?
- ☞ Should regulators, scholars, consumers, and companies assume that soon, AI products may exhibit *agency* over their interactions with the digital and/or the physical world?
- ☞ In the near future, should having *free and open access* to education as well as resources to build or use 'AI-widgets' be considered a civil or a human right?[16]

[14] In general, the level of risk a government is willing to tolerate drive the strictness of regulation.

[15] Critics of this approach point to recurring failures of "big tech" companies in self-regulation. For example, in 2017, Equifax data was hacked and sensitive data including credit history of more than 148 million Americans was stolen. Hackers exploited a known vulnerability in Equifax' software systems to access its database systems. It is reported that the security team at Equifax had failed to fix this issue despite having access to software patch two months prior to the incident (USA Today)

[16] For example, in 2021, the United Nations passed a resolution titled *"The promotion, protection and enjoyment of human rights on the Internet"*.(UN Human Rights Council (2021)).

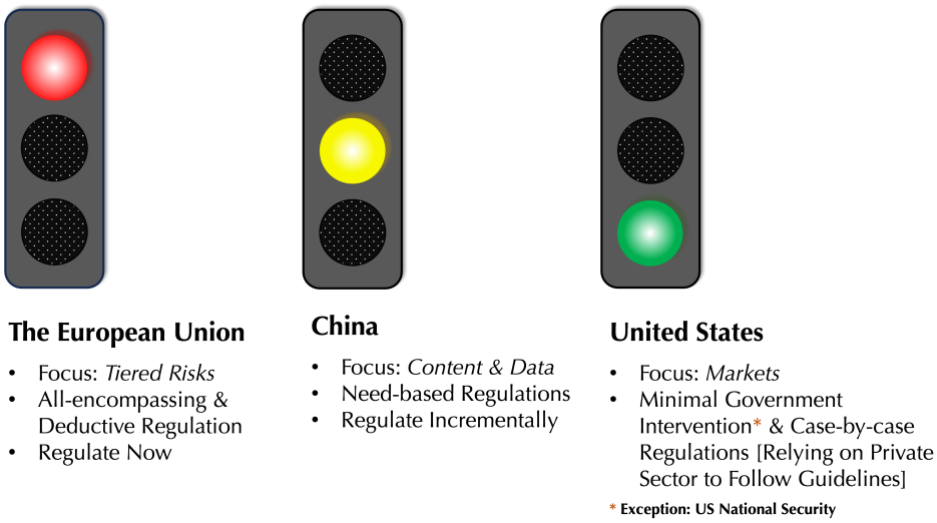

Figure 4: A high-level understanding of EU, USA, and China's legal viewpoint towards regulating AI. Here, **Stop**, **Caution**, or **Go** are various responses to the important question: '*What should one do if existing laws may not have the capacity to regulate AI?*'.

4 AI Regulation: Current Global Landscape

AI regulation is one of the top news topics in the past two years. So how did this all start? We take one step back and like to check the timeline of how different countries responded to online data privacy laws after internet was born. As shown in fig. 3), many countries have very recently passed or enacted online data privacy laws. We argue that a significant technical and legal debt owed to AI regulation is due to challenges associated with enforcing digital data privacy laws.

In this section, we focus on USA, China, and EU's recent announcements and legal activities to regulate AI. Due to the fast pace and rapid development of AI technology, no single country has concluded their AI regulation journey yet. It is easy to recognize different philosophical viewpoints to adopt and regulate AI. In the remainder, provide more details on the latest efforts by every legal entity and their potential implications for the private sector.

4.1 The United States of America: President Biden's Executive Order on 'AI Safety'

On October 30th, 2023, the White House published an executive order titled '*Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*':

> *My Administration places the highest urgency on governing the **development and use of AI safely and responsibly**, and is therefore advancing a coordinated, Federal Government-wide approach to doing so. The rapid speed at which AI capabilities are advancing compels the **United States to lead** in this moment for **the sake of our security, economy, and society**.* (Biden, 2023)

This government-wide EO explicitly directs more than 50 US government entities to devise and implement appropriate actions requested by the White House. More specifically, EO-14110 aims to address 8 overarching policy domains,

- ☞ Safety and Security,

Figure 5: Word Cloud shown above created using the released text of President Biden's executive order titled "*The Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*" (Biden (2023)).

- ☞ Innovation and Competition,
- ☞ Worker Support,
- ☞ AI Bias and Civil Rights,
- ☞ Consumer Protection,
- ☞ Privacy,
- ☞ Federal Government's use of AI,
- ☞ International Leadership,

by directing more than 50 US agencies to adopt and implement specific tasks as well as appropriate guidelines.

4.2 The European Union: EU-AI Act

Commonly known as the '**EU-AI Act**', the European Union recently (December 2023) reached a 'political agreement' on the draft of AI Act. To date, EU-AI Act is the only *horizontal* legal framework towards regulating AI on such scale. At its core, development and deployment of application or services using AI must be categorized from a '*risk management* ' perspective (see figure 6). The five risk categories are as follows:

1. **Unacceptable Risk:** AI systems or products that are assumed to be hazardous to individuals. Examples are:

2. **High Risk:** AI systems or products which could have negative impact on: a) safety; b) fundamental human rights

3. **Limited Risk**: AI systems or products used to create or manipulate contents for human users (e.g. Deepfakes, cf. (Westerlund, 2019)), e.g. audio, video or image.

4. **Minimal Risk:** AI applications such as spam filters and video games are examples of AI applications which pose minimal risk to human user.

5. **'General-purpose Artificial Intelligence' (GPAI):** AI products or systems that are built using 'Foundational Models'[17], cf. Zhou et al. (2023); Schneider et al. (2024). This risk category states that GPAI, inherently, has risk. Amendments further divide this risk into two levels of risk demanding a set of additional requirements[18].

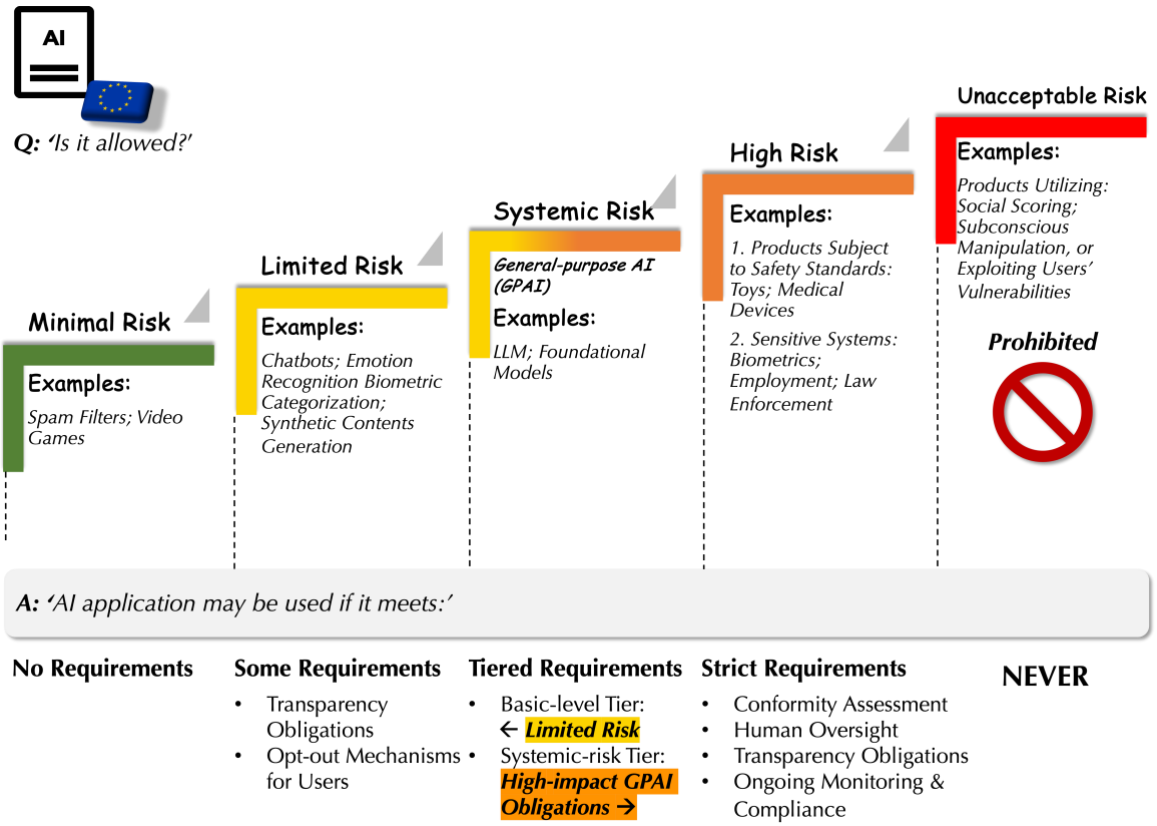

Figure 6: EU-AI Act Risk-based approach towards regulation of AI applications. Any AI-powered application or service is categorized into the five predefined risk categories. Given the assigned category, AI application should follow the set of requirements and legal mandates provided by EU.

[17] First popularized by Stanford Institute for Human-Centered Artificial Intelligence (HAI), a foundation (AI) model is a class of machine learning model (pre)trained to perform a range of tasks with minimal to null tuning effort.

[18] It should be noted that this category was not in the original draft of EU AI Act and was added in a later version in 2023. Mainly, GPAI was added due to rapid emergence of 'Generative AI' products, e.g. ChatGPT or DALL-E.

4.3 China

China has undertaken a hybrid approach towards regulating AI. As China have focused on the most "pressing" and "critical domains" to be regulated first, i.e. social media, online contents, and recommendation engine. In doing so, as of today, three major regulations

1. The Regulation of "**Recommendation Algorithms**". Issued in December 2021 ((Yang and Yao, 2022))
2. The Regulation of "**Synthetic Content**". Issued in November 2022 ((Sheehan, 2023))
3. Interim Measures for the Management of "**Generative Artificial Intelligence**" Services[19]. Issued in July 2023 ((China Law Translate, 2023))

have been implemented. While there is a rich lesson to be learned from China's path towards regulating AI, especially with respect to overcoming technical challenges associated with AI regulation, we remark that China's central government clearly mandates its politically motivated requirements to be at the core of any AI regulation solutions. For example, 'Article 4– Requirement 1' regulation for Generative AI, one reads:

> "...***Content generated through the use of generative AI shall reflect the Socialist Core Values****, and may not contain: subversion of state power; over- turning of the socialist system; incitement of separatism; harm to national unity; propagation of terrorism or extremism; propagation of ethnic hatred or ethnic dis- crimination; violent, obscene, or sexual information; false information; as well as content that may upset economic order or social order.*"

4.4 Other Countries

Almost every country had embarked on 'AI regulation path' before the emergence of powerful 'Generative Artificial Intelligence' (GAI) systems. Prior to GAI, it would seem 'reasonable' to assume there would be ample time for policy makers architect and pass relevant laws. That is not the reality in a post GAI world. Many leaders are now allocating public funding to research and development in TAI and AI safety. Balancing the trade-off between tight-grip regulation and innovation given political, economic, and sovereignty factors is an '*art*'.

1. **UK:** United Kingdom's draft on AI regulation was first released in March 2023. UK's government clearly states a 'pro-innovation' approach towards AI, (UK Government). It states that unlike EU AI Act, UK government would not seek new government units and regulators for TAI.
2. **Japan:** has taken a 'soft' approach towards TAI, i.e. no new regulation has been passed specifically to address TAI. Developers and companies should abide by existing and "closest" laws in data, software, and copyright. In a surprising move, Japan recently announced that use of copyrighted material to train AI models is permitted by law, cf. (ACM News).
3. **Brazil:** Inspired by EU-AI Act, Brazil focuses on a risk-based approach towards regulating AI. In particular, it focuses on the rights of users interacting with AI systems from knowing that they are interacting with an AI agent, demand explanation, or even con- test the decisions made by an AI system, especially for high-risk cases such as financial evaluations, cf. (Holistic AI, 2023).

[19] By many, this law is considered a '*breakthrough*' since it is the first international regulation pertaining to 'Generative AI' technology.

4.5 What can be Learned from China, EU, and USA's Vastly Different Approaches to Regulate AI?

- ☞ EU and China may face similar challenges in balancing trade-off between 'control' and 'innovation'
- ☞ China has taken the lead on drafting the first international regulation of **Generative AI**.
- ☞ While not clear from day one, USA's current path towards AI regulation seems to support making AI openly and widely available, resulting in calls for more support of Open-Source platforms
- ☞ EU's horizontal and deductive view towards AI regulation may seem restrictive. It has been criticized by several member states, e.g. France whose startup industry has been booming on AI and Generative AI.
- ☞ One main benefit of EU's method is that it offers the benefit of longer-term planning and stability for private sector, as frequent updates to the EU-AI Act would not be necessary. Compare this to China's incremental legislation of TAI.
- ☞ In contrast, as for USA and given the precedent-based justice and court system, it is a tedious task to "anticipate" the potential legal shifting landscape via local, state, or federal's perspective.

5 Risk

In March 2022, the Arizona Supreme Court ruled that ((Supreme Court of the State of Arizona, 2022)) the family of a 4-year-old girl named *Vivian Varela*, who had been killed in 2015 in a car accident, can sue Fiat Chrysler Automobiles, the parent company of Jeep, for '**wrongful death**'. The family had argued that the automatic emergency braking system, which could have potentially prevented the crash, was not installed in the 2014-Jeep Grand Cherokee that rear-ended their car. Despite the availability of this life-saving technology[20], at the time, it was only offered as an optional feature bundled with a "*luxury package*" for an additional $10,000. In hindsight, the fatal crash could have been prevented if companies prioritized safety over profits. Jeep's decision to treat the '*emergency braking system*' as a financial incentive rather than a standard safety feature reflects a misguided approach[21].

AI-enabled decision-making tools, sometimes referred to as 'digital twins', are becoming integral parts to various fields including engineering, business, human resources, procurement, and government. As their use continues to proliferate, we expect an escalation in the complexities surrounding ethics, engineering, and profitability. In extreme instances, the legal implications may thrust any court/justice system into uncharted territory, potentially, establishing new legal precedents.

[20] Multiple studies have reported 40% to 70% fewer rear-end and front end crashes, cf. (Cicchino, 2018), (Aukema et al., 2023), and (Fildes et al., 2015)

[21] In 2014, installing an emergency braking system was not a governmental mandate. Therefore, Jeep treated it as a 'luxury' feature ought to be purchased by customers.

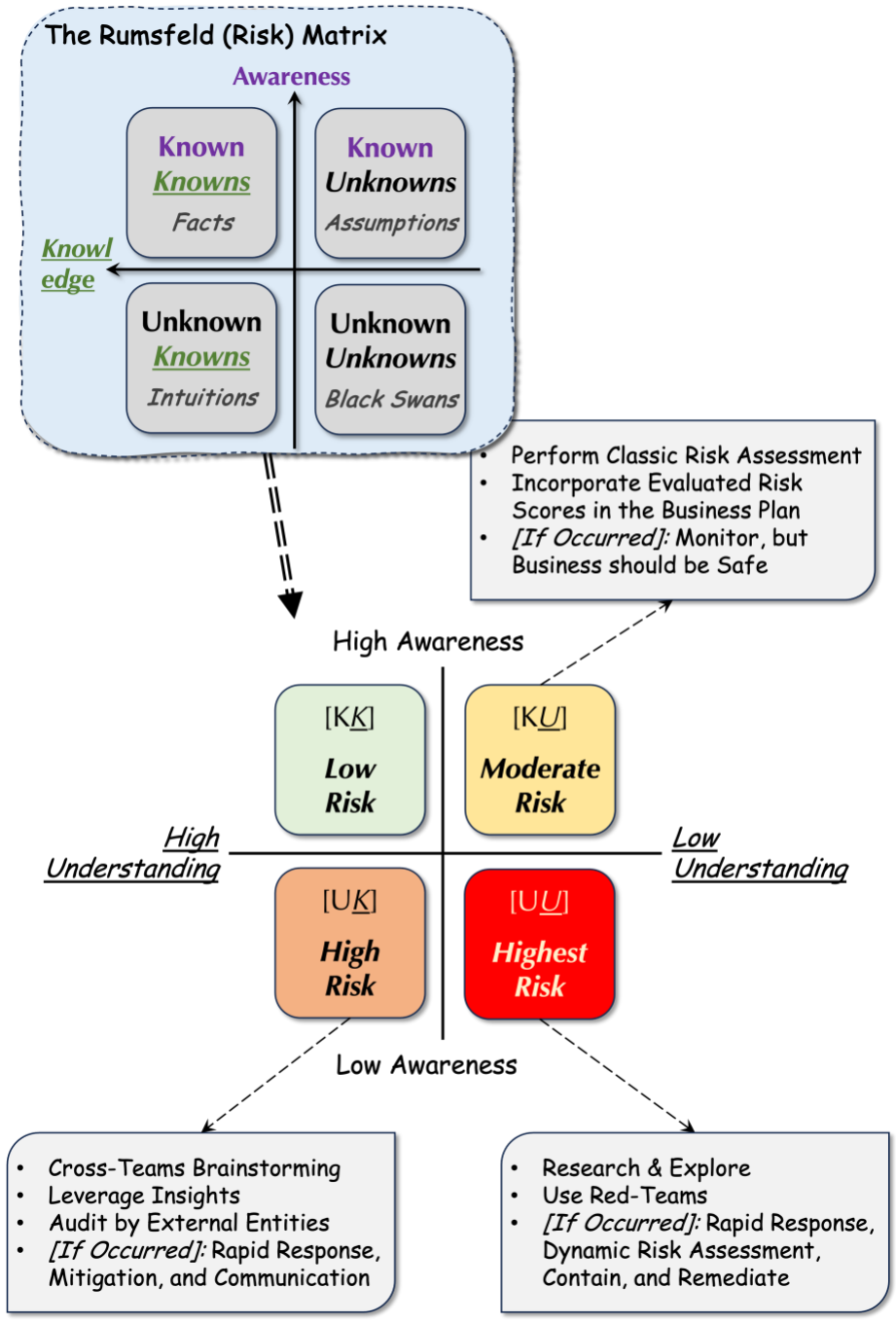

Figure 7: Risk quadrants (also known as the Rumsfeld Risk Matrix (RRM)) and common recommended action for each risk level. Here, U*U*, U*K*, K*U*, and K*K* refer to Unknown *Unknown*, Unknown *Known*, Known *Unknown*, and Known *Known* respectively. It is important to consider the action plans to mitigate risk according to each region. RM can be employed by any team building or utilizing an AI system to plan for, mitigate, or remediate potential risks or legal challenges.

5.1 Managing Risk and Making *Good* Decisions under Uncertainty

Within any organization, managers and decision makers are expected to understand, plan, mitigate, and navigate risks. Disciplines such as 'Operations Research', 'Enterprise Risk Management' (ERM) (cf. Bromiley et al. (2015)), 'Strategic Management', are only a few examples. In general, such disciplines aim to combine structured, empirical, and statistical frameworks so that managers facing uncertainty, could plan for risks or make informed decisions. Often, uncertainty is rooted in having an incomplete view/data into the status of company, product, demand, clients, customer behavior, or true randomness, also known as '***aleatory uncertainty***'[22].

In the context of TAI, it is important to recognize how every category of uncertainty can be estimated, measured, detected, reduced, or eliminated. Furthermore, transforming such 'unknowns' into 'risk score' or 'risk level' compatible with existing ERM is a non-trivial task. While ERM for IT and Cybersecurity has been a well-studied discipline, to the best of our knowledge, there is no widely-accepted framework for how to incorporate TAI to ERM for every organization.

As the first step, with the adoption of UQ by the AI scientific community, we now have access to several mathematical and statistical techniques estimating uncertainties associated with an AI model output (cf., Hüllermeier and Waegeman (2021); Gawlikowski et al. (2021)).

As an example, consider a car with an 'intelligent' *automatic brake system* which uses computer vision to detect nearby objects and preventing collisions. This module is, however, designed to operate to assist the human driver only in 'normal or acceptable' conditions. In an event where driver is facing unfavorable conditions such as extreme fog, this intelligent system must be "self-aware" of its lack of '***confidence***' in the outputs returned by the computer vision module and used to warn the drive and disengage.

In this chapter, we select the 'Rumsfeld Risk Matrix' (RRM) and apply risk management in the context of AI and TAI. In doing so, we demonstrate how a simple framework such as RRM be incorporated into AI products or systems and map the risk categories associated with every step of an AI product life-cycle into 'actionable' insight required for efficient implementation of TAI.

5.2 Example: Collecting Training Data and Mapping Risk to Actions

In the context of **collecting training data** to build a new AI product, following are scenario examples of risks and how they could fall under each quadrant in the RM matrix.

(i) **Known *Known*:** Collecting user activity data– training data– from a *biased* source, such as a social media platform where users are more likely to express extreme views.

- ☞ Source of training data is known to be unreliable or biased.
- ☞ Training data is inaccurate or incomplete.
- ☞ Training data includes sensitive information that could be used to discriminate against certain groups of people.

Mitigation strategy: Diversify data sources to reduce the risk of bias.

[22] Aleatory uncertainty refers to inherent randomness in a phenomenon that can never be predicted, e.g. outcome of a (fair) coin toss. In contrast, '*epistemic uncertainty*' (also known as '*systematic uncertainty*') refers to inaccuracies in data or observations that can be reduced or eliminated by means of more experiments or collecting new data. For a review on aleatory and epistemic uncertainty, cf. Der Kiureghian and Ditlevsen (2009).

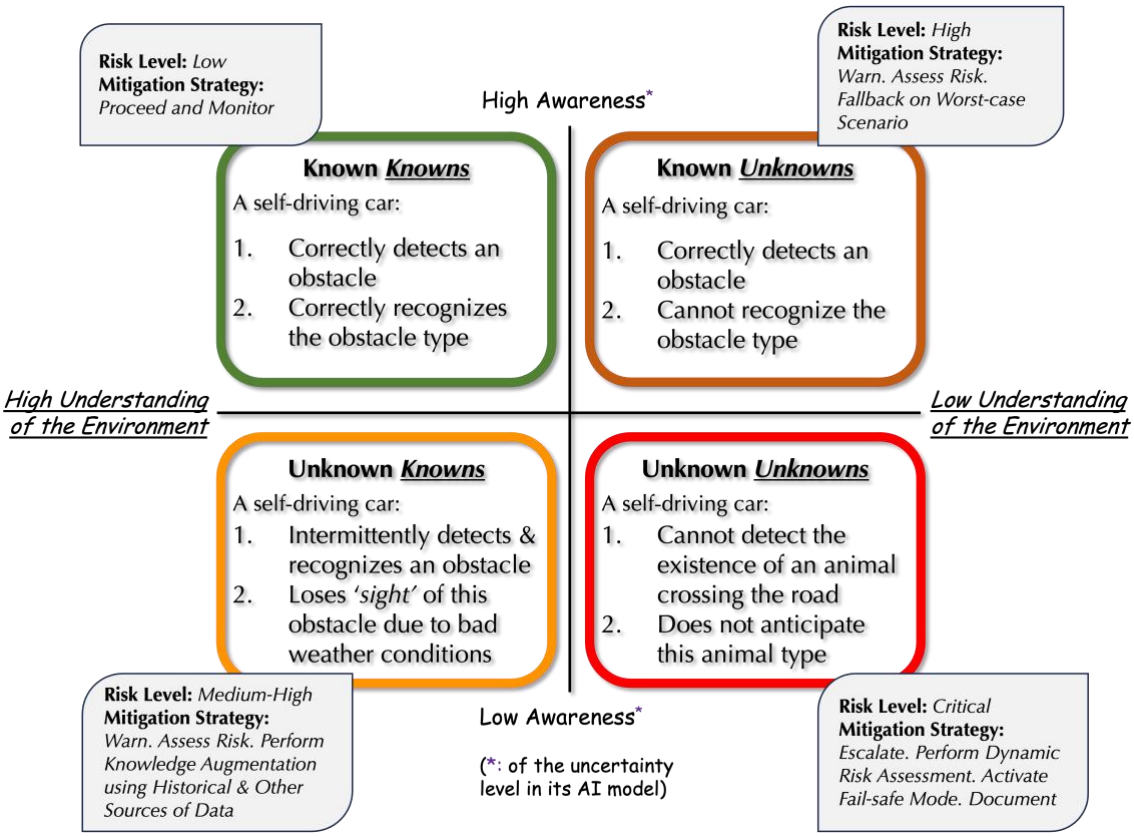

Figure 8: Rumsfeld Risk Matrix (RMM) constructed for a hypothetical scenario: An AI model utilized by a 'self-driving car' to identify objects, humans, and animals on the road. This is an example of how using RMF during or after training and *productionizing* an AI model system can benefit the engineering or test teams. Depending on the risk level, quadrant, managers, and stakeholders can estimate the risk associated with every quadrant, and adjust the AI model or mitigation resources, accordingly.

(ii) **Known *Unknown*:** Collecting training data that is not perfectly accurate, but is still sufficient for building an AI model.

- ☞ Collecting training data from a new or emerging source that has not been previously evaluated for quality or bias.
- ☞ Collecting training data from a source that is known to be reliable, but the specific data being collected has not been evaluated for quality or bias.
- ☞ Collecting training data that is known to be accurate and complete, but the potential impact of using the data to train an ML model is unknown.

Mitigation strategy: Apply data quality control measures to identify and correct– if possible– errors, e.g. missing data, in the data.

(iii) **Unknown *Known*:** Collecting training data that is highly sensitive and if not handled by experienced data scientists, could result in discrimination against certain groups of people.

- ☞ Collecting training data from a source that is unknown, but the likelihood of the data being biased or inaccurate is known to be high.
- ☞ Collecting training data from a source that is unknown, but the potential impact of using the data to train an ML model is known to be high.

Mitigation strategy: Implement strong data security measures to protect the data. This category of data can be only used based on per case-by-case and approval by 'Chief Information Officer' (CIO)

(iv) **Unknown *Unknown*:** Collecting training data that that does not include its meta- data or proper documentation on its original source. Therefore, it is not clear if this dataset is particularly relevant to the task which the new AI model will be used for.

- ☞ Collecting training data from a source that is completely unknown, and both the likelihood of the data being biased or inaccurate and the potential impact of using the data to train an ML model are unknown.

Mitigation strategy: Investigate further with other teams to find out the source of data. If applicable, conduct exploratory analysis in a protected and 'sandbox' environment.

5.3 AI Regulatory Sandbox: A Useful and Interim Medium

We firmly believe that we should use all the means to allow innovation in the AI domain alive. Provisions mentioned in the EU-AI Act and WHEO– with reasonable intentions– could ultimately stifle innovation as well as engagement at the community levels. We have yet to observe the actual implementation and guidelines– as they say, the devil is in the details[23]

To mitigate this, EU-AI act introduces a new concept called '**AI Regulatory Sandbox**' which encourages the EU members to create regulatory environments, tools, and best practices for testing and experimentation with new AI products– under supervision of EU members and approved authorities, cf. (Truby et al., 2022). In essence AI regulatory sandbox serves two purposes:

[23] If a few of such provisions are not implemented tactfully, we believe that it could lead to a state where only a few wealthy and resourceful conglomerates could "afford" the risks and subsequent legal fines provisioned in EU-AI Act. In other words, individuals and startups driving any meaningful innovation in TAI are discouraged.

I. Foster learning and innovation in AI for businesses via real-world development and testing of new AI-powered products.

II. Contribute to regulatory learning by creation and testing experimental legal frameworks around new technologies based on AI.

While this provision in EU-AI act has yet to be finalized, Spain has recently launched the first program of this kind to foster AI innovation while evaluating regulatory requirements to be enacted in EU-AI Act. This point of view seems to be gaining a widespread interest as it aims to expand beyond EU. For instance, Sam Altman– CEO of OpenAI – recently invited the 'United Arab Emirates' (UAE) to become a testing ground for AI regulation (Bloomberg (2024)).

6 Bias and Fairness

Key Takeaways

- ❒ There are multiple definitions for 'fairness'.
- ❒ Mathematically speaking, it is proved that not all aspects of fairness– characterized by definitions– be enforced concurrently.

6.1 'Biased AI': A Polysemic Term Which Needs Clarification

For better or worse, diverse community surrounding AI have been using the term '**biased**' AI to often disparate technical or conceptual topics. This may have caused unnecessary ambiguity and sometimes confusion, cf. Felzmann et al. (2019). To 'decode' this term, it is important to pause and reflect on two main clarifying items with respect to any biased AI system: 1) What is the context that AI product is used? and 2) Who is the SME and his/her role in AI life cycle? For instance, consider the following SMEs studying or mitigating bias in AI-enabled system:

1. **AI Engineer/Data Scientist:** Algorithmic bias – The systematic error introduced by the design and implementation of machine learning algorithms. While an entirely mathematical concept, if not detected properly, it may result in unreliable or even unfair outcomes, cf. Barocas and Selbst (2016); Kordzadeh and Ghasemaghaei (2022); Belkin et al. (2019); Curth et al. (2024).

2. **Regulator/Policy Maker:** Social or human bias – The unfair and prejudicial treatment of certain individuals or minority groups usually caused by pre-existing societal and historical biases reflected in the data used to train AI models, cf. Buolamwini and Gebru (2018); Noseworthy et al. (2020).

3. **Ethicist/Philosopher**: Ethical bias – The moral implications of AI decision-making, which may involve value judgments, unequal treatment, or perpetuating existing social inequalities, cf. Hagendorff (2022); Jobin et al. (2019); Mittelstadt et al. (2016).

4. **Data Analyst:** Statistical bias – The difference between an algorithm's expected prediction and the true value, which can result from errors in data collection, sampling, or modeling assumptions, cf. Hastie et al. (2009).

5. **User Experience (UX) Designer:** Interaction bias – The biases that emerge from the design of AI interfaces and how users interact with them, potentially leading to

unintended consequences or unequal access to AI-driven services, cf. Yoon and Jun (2023); Bach et al. (2022); Meske and Bunde (2020); Oh et al. (2018).

6. **Social Scientist:** Systemic bias – The ways in which AI systems can perpetuate and amplify broader social, economic, and political inequalities, cf. Beer (2019); Fountain (2022)

7. **Legal Scholar:** Legal bias – The potential for AI systems to generate outcomes that violate existing laws, regulations, or legal principles, such as those related to nondiscrimination, privacy, or even due process[24], cf. Citron and Pasquale (2014).

These interpretations demonstrate the *polysemic* nature of the term 'Bias' in AI and ML, as its meaning can vary significantly depending on the context and the persona using it.

6.2 Bias as State-of-mind of an Individual

"*Quis custodiet ipsos custodes? (Who will watch the watchmen?)"* (Decimus J. Juvenalis (circa 1st/2nd century CE); A Roman poet)

Human oversight is paramount in the end-to-end life-cycle of AI models. This also includes auditing and monitoring systems designed to deliver a medium for TAI. However, with humans within any organization, we anticipate that their decision-making process is not immune to multiple forms of cognitive bias in human brain. Scholars in psychology have conducted an exhaustive search and indicated how this bias kind could overshadow honest interpretation of any situation. More recently, there have been studies to harvest this knowledge from the field of psychology to tackle various forms of biases hurting AI systems, cf. Tambe et al. (2019); Ashmore et al. (2021). Without going into details, below is a a few common categories of biases known to impair human judgement:

- ☞ Cognitive Bias
- ☞ Confirmation Bias
- ☞ Anchoring Bias
- ☞ Ethical Fading
- ☞ Primacy Effect
- ☞ Group-think and Conformity Paradox
- ☞ Self-serving Bias
- ☞ Moral Licensing

We note that understanding how the list above can impact SMEs in charge of overseeing or investigation potential problems within data, AI model, audits, testing, or quality assurance is a must. considering above list in drawing conclusion is key. For every bias type, and in the context of TAI, there are different mitigation strategies that can help decision makers and SMEs minimize the risk imposed by cognitive bias in the handling of TAI system.

[24] Legal scholars would be particularly interested in understanding how AI systems can be designed, implemented, and governed to ensure compliance with existing law and protect individuals' rights. They could also consider the challenges of holding AI systems and their creators accountable for biased outcomes, auditing AI systems without violating intellectual property rights, and the potential need for new legal frameworks to address these issues.

For example, in their investigation, De Fuentes and Porcuna (2019) show that financial risk assessment reports conducted by various independent entities in Europe seemed to be depended on:

1. The magnitude and impact of the financial catastrophe or scenario under review.
2. The total number of signatories of produced report, i.e. only one individual (*vs* more than one person signed the final report.

De Fuentes and Porcuna (2019) conclude that the auditing firms seemed to be more concerned about the *public reaction* to company's reputation and, therefore, tried to avoid any potential media scandals because of their findings shared in public reports. While it may seem dire, in the context of TAI, there has been systematic approaches that can minimize or eliminate biased decision of the '*watchmen*'. It is beyond the scope of current paper to dive into such mitigation strategies. For more insight on this topic, we refer reader to Wall et al. (2017); Mohanani et al. (2018); Kliegr et al. (2021) and references therein.

6.3 Fairness

There seems to be rather universally accepted 'conventional wisdom' states that any *mathematical* definition of '*fairness*' is often at odds with *human perception* of fairness. Mathematical notion of fairness would focus on measurements (data) as well as clearly defined equations accompanied by the proper statistical framework to be used. On the other hand, human perception of fairness, tends to be '*descriptive*', i.e. a person may verbally share that she "felt" being treated unfairly by an agent. In doing so, she would commonly employ relativistic or contrastive arguments to prove her point (Srivastava et al. (2019)).

For example, consider a hypothetical scenario where a passenger at the airport is asking to be upgraded to the 'business class'. Upon checking with the gate agent, her request is turned down. In her complaint submitted to the airline customer service, passenger states the following as ground for being *unfairly treated*:

> "*I have been a loyal customer of Almost-Landing Airlines*[25] *for 10 years with a frequent flyer status.* ***I was denied an upgrade*** *to the business class, even though there were plenty of available seats. Also, I witnessed* ***another passenger without a frequent flyer status being upgraded*** *without any issues. That is* ***not fair****...*"

This example shows an individual perception of unfair treatment which may never exist in 50 years from today. As human society advances, so do new definitions– read perception– of 'fairness' (for an overview of how fairness metrics evolved in the past 50 years, cf. cf. Hutchinson and Mitchell (2019)).

6.4 Widely Accepted Definitions for Fairness

There is no unique metric which defines fairness. Any definition of fairness which is accepted by society[26], has yet to be transformed into mathematical or statistical manifestation. Once this is agreed up, any AI-system (or a statistical inference module) can be subjected to a 'fairness assessment' compute engine which in turn would quantify a 'fairness score'. This

[25] We trust that our reader infers such names are completely fake and do not represent any particular airline.

[26] We remark that the notion of fairness not only depends on the existing cultural context, but it can also vary over long periods of time, cf. (Saxena et al., 2019). Consider how recently women were allowed to vote, even in first-world countries such as the US or Switzerland

would be the first critical step to remediate any unintentional and unacceptable unfair decisions made by an AI-system. Alternatively, one can directly '*map*' computed fairness score to a risk score (e.g. for the case of EU-AI-Act) and plan accordingly. Below is a widely accepted list of fairness metrics used by scholars in TAI research (collected from (Mehrabi et al., 2021) and references therein):

- Equalized Odds (Group)
- Equal Opportunity (Group)
- Demographic Parity (Group)
- Fairness Through Awareness (Group)
- Fairness Through Unawareness (Group)
- Treatment Equality (Group)
- Test Fairness (Group)
- Subgroup Fairness (Subgroup)
- Counterfactual Fairness (Individual)
- Fairness in Relational Domains (Individual)
- Conditional Statistical Parity (Individual)

6.5 Fairness Through the Lens of Group Size

According to the number of humans impacted by the outcome of an AI model there are three classes of enforcement:

1. **Individual Fairness:** Aims for similar predictions (produced by AI model) to "similar" individuals. For example, assuming a male professor and a female professor with similar financial backgrounds apply for a new 'credit card', they both should receive similar maximum credit line. In essence, in a *fair* system, gender is considered a protected attribute and should not determine *credit worthiness* of an applicant.
2. **Group Fairness:** In a large population, various *groups* characterized by 'sensitive' or 'legally-protected' indicators, e.g. race or gender[27], should be treated '*equally*'.
3. **Subgroup Fairness:** Combines both viewpoints above to have a better outcome. For example, consider an AI model that screens candidates' application for an on-site interview. If an average '*false negative rate*', i.e. '*Applicant Rejected* ', for **female** applicants– a subgroup defined by the attribute gender– is significantly higher than that of **male** applicants, gender-aware subgroup fairness may have been violated (for an indepth discussion, see (Kearns et al., 2018)).

To make matters even more challenging, there has been mathematical proof indicating that enforcing all definitions of fairness simultaneously is not possible, cf. (Kleinberg et al., 2016).

For example, a bank could attempt to be fair to all its clients by enforcing 'fair' AI in one category, by would ultimately be 'unfair' to another group of customers.

[27] A group may be defined differently considering legal or business requirements. For instance, according to legal mandates outlined in Title VII of Civil Rights Act of 1964 (act (1964)), any AI model involved in hiring decisions should be designed such that race, color, religion, sex, or national origin does not play a decisive factor in employment.

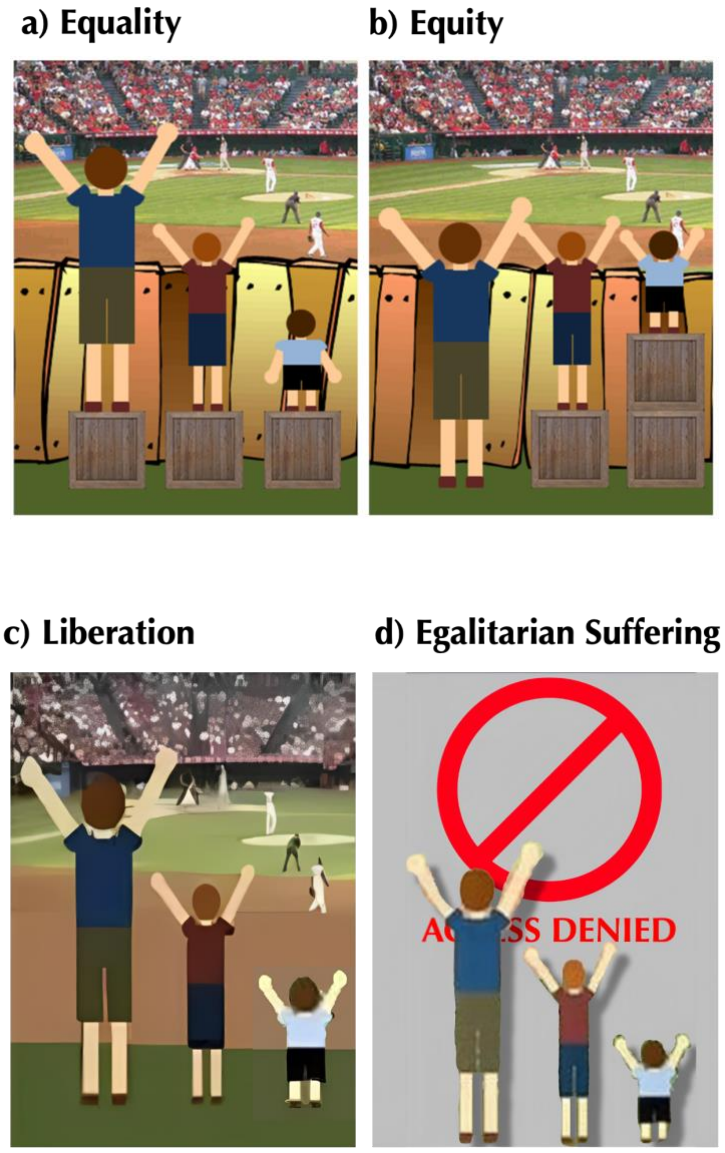

Figure 9: A simplified depiction of different definitions of 'fairness' and associated interventions aimed to implement 'justice'. (For a full story on the history of this meme, see (Froehle, 2016))

6.6 AI Fairness and Human Rights: COMPAS Example

'Correctional Offender Management Profiling for Alternative Sanctions' (COMPAS) built by a private company called Equivant has been used by U.S. courts to assess the likelihood of a defendant becoming a '*recidivist*' in the next two years. Judges in the past have used this product to make decisions on defendant requests submitted to courts. While this topic has been widely debated amongst independent investigators, it is imperative to note how having limited to no-access to inner-workings, training data, as well as the audit results of a software like COMPAS have eroded the public trust. Below is a list of concerns raised by independent scholars:

- ☞ **Human Rights:** COMPAS is a propriety product built by a for-profit company. Currently, it is closed to public scrutiny. Not only this violates 14th Amendment right, but COMPAS' lack of transparency has caused conflicting conclusions carried out by independent investigators, cf. Rudin et al. (2020).
- ☞ **Human *vs* COMPAS:** In a recent study, it is shown that COMPAS does not perform better than an average conclusion derived from a pool of random strangers who are also not familiar with the criminal court systems, (Dressel and Farid, 2018)[28].
- ☞ **Racial Bias:** COMPAS has caused controversy as it may violate '14th Amendment Equal Protection' rights on the basis of race[29], since the algorithms are argued to be racially discriminatory with disparate treatment of African-American defendants, cf. (Thomas and Ponton-Nunez, 2022).

6.7 Our Proposed Solution: Example Template for 'Fairness Verification and Validation Testing'

Let's assume a business has an overarching team, 'Fairness Verification and Validation Testing' (FVVT) who is responsible for the enforcement of fairness policy by product teams using AI models. FVVT team proceeds to engage and collaborate with several SMEs to:

- ☞ **Comprehending Legal Requirements:** Understand and manifest legal mandates and associated risks for business, customers, or other entities.
- ☞ **Identifying Impacted Product Teams:** If not clear, consider including every engineering team involved in data collection, training, deployment, and monitoring of an AI model.
- ☞ **Operationalizing Legal Requirements**: a) Define or choose existing (Mathematical/Statistical) Fairness Metrics and KPIs. b) Map legal requirements into 'acceptance criteria'.
- ☞ **Select/Build Benchmarks:** Identifying relevant *test* scenarios considering designated fairness dataset, AI model, and compute resources.
- ☞ **Test and Report:** Run agreed upon FVVT-benchmarks, report test results and metrics used, extract insights and report to various teams involved. If any violations ob- served, understand if current enterprise risk management strategy can remediate such violations.

[28] In their study, authors recruited 400 volunteers. Every person was then asked to guess whether a *defendant* would commit a crime within two years after studying a summary on defendants published by ProPublica.

[29] Significant disparities in the recommendations returned by COMPAS software have been reported for *African-American* and *Caucasian* defendants (Dressel and Farid (2018)).

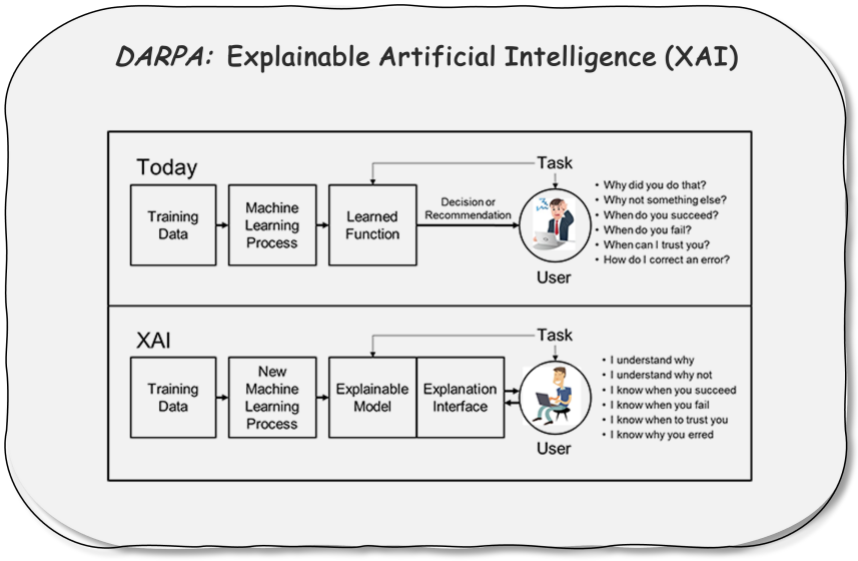

Figure 10: Schematics of XAI first introduced by DARPA in 2017, Gunning (2017).

☞ **Governance:** Versioning and document tests, testing dataset, selected benchmarks, human-readable interpretation of test results, AI models meta data, risk level, fairness KPIs.

7 Explainable AI as an Enabler of Trustworthy AI

First coined by David Gunning[30] in 2017 (see fig. 10), the term 'Explainable AI' has become a priority for research labs and companies. In DARPA's initial framework, XAI was to build machine learning and mathematical techniques which can be used to "comprehend" outputs produced by any black-box AI models, e.g. DNN. By doing so, an end user could rationalize 'outputs' produced by an AI model before making (business) decisions.
Over a short period of time, however, the expectations for XAI have significantly expanded in terms of diversity of end users as well as the level of details in XAI reports. With the expansion in scope, it is not surprising to see XAI research or taxonomy overlapping with other features of TAI.

7.1 XAI: Spectrum of Explainability and Interpretability

More adoption and 'infusion' of AI-systems within Enterprise business cycles impose higher stakes for different end users. We argue that XAI tools and framework focused mostly on the realization of XAI for AI/ML experts. However, with the expansion of XAI to other stakeholders, more strict government regulations, and significant growth of black-box AI models, XAI is now different.

Any organization implementing XAI in their business should expect that XAI ought to produce reports should consider main inputs: WHO, WHEN, and WHY. In other words, 'explainability' component in XAI is dependent on the answers above. Hence, various stakeholders in an organization should expect a 'spectrum of explainability' be produced. This

[30] David Gunning who is currently retired was a program manager at 'Defense Advanced Research Projects Agency' (DARPA).

spectrum would have various levels of details, e.g. statistical terms to plain text, granularity, or even stratified level of access to the data or organizational chart.

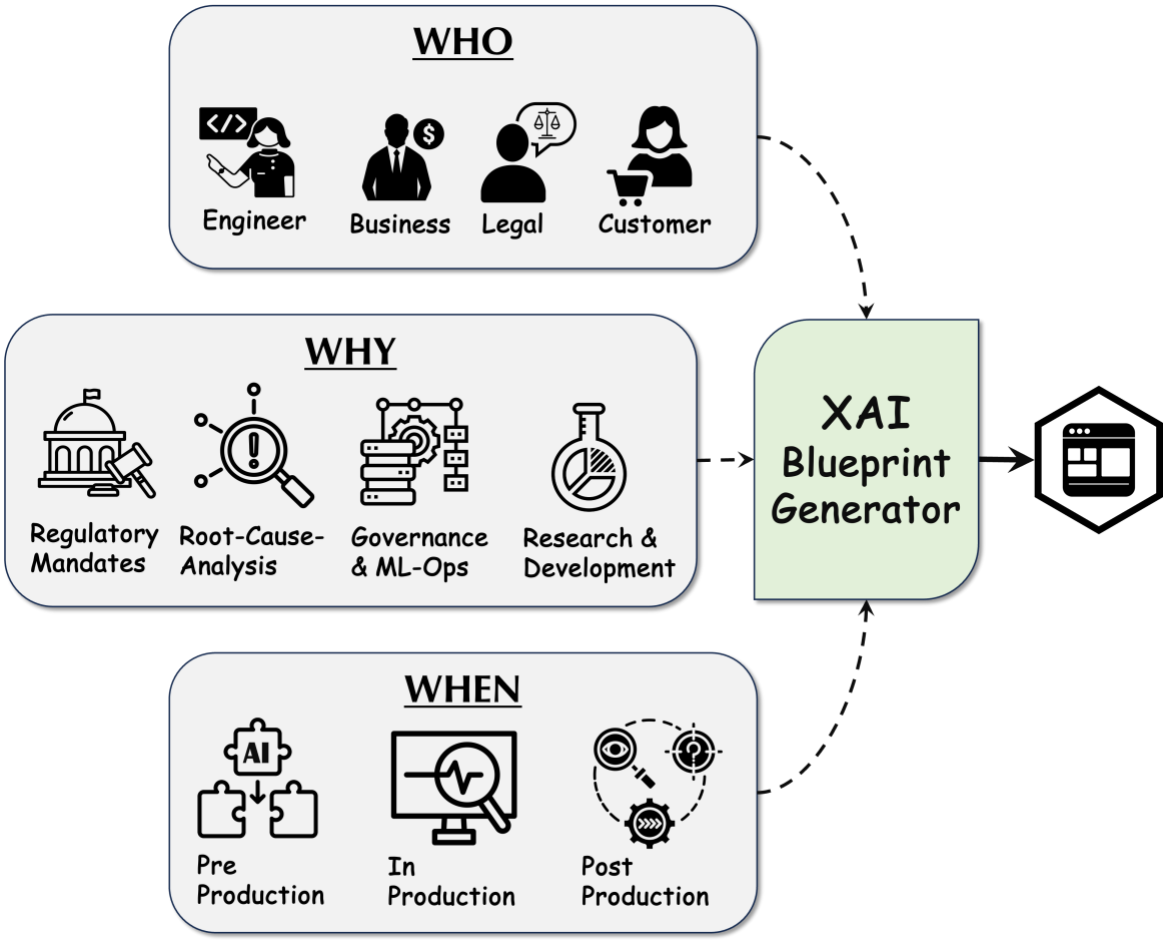

Figure 11: Our proposed framework using three main inputs to generate an 'XAI Blueprint' for an organization. For an example, see table 2 in text.

7.2 Our Proposed Solution: XAI Blueprint Generation

Any enterprise or entity that needs XAI (as an independent or part of any TAI solution), ought to ask three questions. The sheer number of **feasible answers** to the following questions (see fig. 11),

- ☞ ***Who*** *needs XAI?*
- ☞ ***Why*** *is XAI needed?*
- ☞ ***When*** *is XAI needed?*

can easily overwhelm a midsize to large organization, and, therefore, could render the adoption of XAI impossible. In our solution, we propose a flexible framework which answers above would be passed to an XAI Blue Generation engine to create an appropriate actions, setups, and requirements to enable XAI. In table 2, we demonstrate this via three different end users and AI product at three different stages mentioned above.

Table 2: Generating an 'XAI Blueprint' based on our proposed framework. We propose prioritizing answers to "Who, When, and Why XAI is needed?" to product a relevant XAI Blueprint with relevant questions to be addressed for an efficient implementation of XAI.

Who?	**When?**	**Why?**	**A Sample Generated XAI Blueprint**
AI Engineer	Pre-production	Help choose the best AI Model Type so that it Passes the New Regulatory Mandates by EU.	☞ Does AI Model Outputs Classification or Regression? ☞ Is Uncertainty Quantification Needed? ☞ Do Legal Mandates Require Non-expert Reasoning Logic? ☞ Training Data should Exclude Sensitive Features
Legal Team	In-production	Response to a Customer's Inquiry: '*Why was my Loan Application Denied?*	☞ What is the Identified Risk Level using EI-AI Act? ☞ Governance: Which AI Model was used to Make Decision for Current Applicant? ☞ If a Black-box AI Model, Estimate how much Additional Time is Required to Produce the XAI Report. ☞ What are the Mitigation Plans if AI Model shows any Legal Violations?
Business Analyst	Post-production	RCA: Why Loan Default Rates has Increased in the Past Quarter	☞ Governance: Business-friendly Features ☞ Compare against Historical Trends. ☞ If AI Model was Updated or Retrained, which KPIs were Used to Validate the Updates it? ☞ Examine if the AI Model is still Relevant: Has Concept or Data Drift Occurred?

8 Implementation Framework

In the remainder, we provide a few examples of how such frameworks are needed to implement, monitor, and enforce TAI. One point to consider is that adopting one framework is not mutually exclusive. We claim that, more often than not, depending on the context, different frameworks have to be implemented in order to facilitate integration of 'trustworthiness' in an AI-system. For example, if a private company uses an AI tool to help its employees for internal use, it may use a less stringent TAI framework[31] as opposed to an AI-platform sold to external clients. Each of these entities need their own TAI framework, respectively. We remark that it is reasonable that such frameworks are different since each party has its own goal, legal requirements, or resources altogether.

8.1 Trustworthy-By-Design

Given 'trustworthiness' (or any other characteristics of TAI) is enforced before or during the designing step of product development, any attribute associated with TAI is translated as additional constraints. These constraints may restrict, *a priori*, the choice of AI model type, training and/or testing procedure, acceptance criteria, and set of objective functions. On the bright side, such seemingly stringent conditions would help build an AI product that is– at least to a good degree– '*trustworthy-by-design*'[32].

Let us use a simple familiar example. Consider a car manufacturing company about to build and release a new model. For the sake of example, let's consider the following scenarios concerning legal mandates on speed-limit violations (see fig. 12):

- ☞ **Scenario A:** Driver is liable for any speed-limit violations.
- ☞ **Scenario B:** While driver is still liable for any speed limit violations, car companies are now required (by law) to warn drivers when surpassing the speed limit of 65 mph.
- ☞ **Scenario C:** New legislation holds any car company liable for any speed-limit violations. Lawmakers expect that the new car models *by design* could not support speeds above 65 mph.

We hope that the example of a new car design choice by legal and engineering team states the challenges with defining a 'trustworthy-by-design' framework when building an AI product. In scenario C, engineering team decides to design the car engine capacity such that it will not be capable of passing the speed limit.

8.1.1 Need-to-Know-Basis

Despite their remarkable performance, applications built using LLMs or GPTs, have demonstrated weak to no strains against revealing 'too much' information (Greshake et al. (2023)). Several examples of how human users 'design prompts'[33] (or inputs) to systematically extract sensitive information from AI models (Liu et al. (2023); Yao et al. (2024)). For example, a 13-year-old student interacting with an AI tutoring software, should not be provided information on how to use drugs– at least without the supervision of teacher in the AI-student

[31] To the extent required by law.

[32] It is sometimes helpful to consider example products in the *physical world*. In contrast to digital products (e.g. Email), products in physical world provide a more intuitive sense of 'rights', 'legal mandates', 'reliability', or 'accountability'.

[33] Also known as 'prompt injection'.

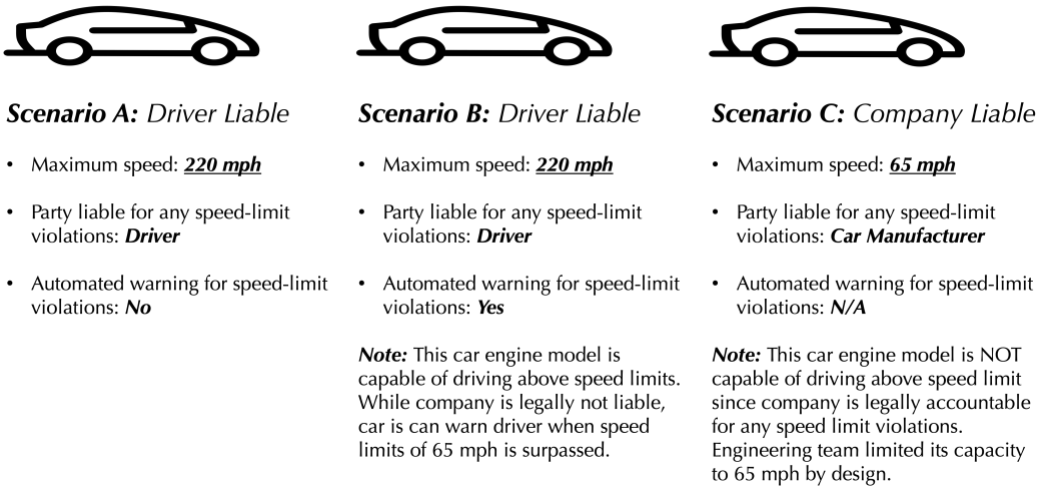

Figure 12: Illustration depicting how three hypothetical scenarios for legal liabilities can impact the design choice and engineering of any new AI product. We are using the example of a new car being manufactured to simplify the topic of 'trustworthy-by-design' for an AI system. In scenario C, engineering team decides to design the car engine capacity such that it will not be capable of passing the speed limit.

engagement session. With no universal solution readily available, concerns for user safety, privacy, IP theft, or fraud using 'adversarial attacks' on AI models is an active research topic, cf. Qiu et al. (2019).

In essence, an AI-powered system should only '*know*' the '*knowledge*' needs to fulfill the task(s) it is built for. For example, one strategy is to ensure that training data (along with its relevant features or meta-data) that is intended to be used and build an AI product, must not contain information that are not relevant to the use-case. Overall, mitigation strategies in making AI products resilient against adversarial attacks and prompt injection depend on factors such as compute resources, risk level, and legal constraints, cf. (Rai et al., 2024).

8.2 Trustworthy Assurance

Unlike 'trustworthy-by-design' (see section 8.1), 'trustworthy assurance' aims to test and verify trustworthiness (or a subset of attributes of TAI) '*after*' creation of any (AI) product or services. Thus, making decisions at the design and/or training step to ensure 'trustworthiness' may not be required. In rare cases, an AI model can first be *trained* while fully disregarding any TAI attributes. Next, 'Trustworthy Assurance' team proceeds to conduct its family of predefined tests for required attributes of TAI. If any of those tests failed, results would be reflected in meaningful reports and can be handed to the engineering and legal teams for potential updates or corrections to the original model. While this approach has been adopted by many companies for the past few years, we do not recommend this as many countries are anticipated to pass laws favoring or mandating 'trustworthy-by-design' methodology for implementation of TAI in the private sector.

8.3 Trustworthy via Continuous Monitoring and Improvement

As the title suggests, this framework is not exclusive from the ones discussed in 8.1 and 8.2. Accepting '*The only constant in life is change*'[34], any AI model/product–regardless of its initial state– should be continuously monitored, tested, and improved. Reasons for adhering to this philosophy goes above ensuring 'trustworthiness'. Continuous monitoring and improvement life-cycle has been applied to traditional software development for years (for a survey on approaches and practices, cf. (Shahin et al., 2017), (Mishra and Otaiwi, 2020)). More recently, similar frameworks such as 'Machine Learning Operations' (ML-Ops) introduced to facilitate challenges associated with continuous development, monitoring and deployment of AI/ML models in the production settings for any business (cf., (Kreuzberger et al., 2023), (Symeonidis et al., 2022), and (Testi et al., 2022)).

8.4 Our Proposed Solution

We propose a risk-based and flexible framework to help enterprise implement as well as operationalize TAI in their business. The framework has four distinct steps (see fig. 13):

1. **Set:** Goal is to clearly set 'Legal & Policy Requirements' (LPR) along with 'Risk Levels and Thresholds' (RLT) that must be considered with the AI-system.
2. **Formalize:** LPR and RLT to define or select TAI-aware measurable metrics, KPIs and acceptance criteria. In addition, formulate proper equation to set up benchmarks.
3. **Measure:** Upon running benchmarks or measuring the metrics from Formalize step, record potential violation flags based on configured acceptance criteria.
4. **Act:** Interpret findings, metrics, and flags from previous step back to non-technical (if needed) implications. If necessary, escalate or suggestion mitigation plans to minimize risk (of violation).

9 A Few Suggestions for a Viable Path Forward

9.1 Continue Supporting Academic Research in Trustworthy AI

'Center for Security and Emerging Technology' (CSET)[35] analyzed prior scientific publications on topics related to TAI. Using a thorough and systematic analysis to contextualize the trustworthy AI terms in more than 30,000 scientific publications, CSET identified 18 clusters relevant to publications in set or subset of TAI, (Toney and Probasco, 2023). The list of top publishers including only one private company that made it to the list (Google) is given in table 3 (source: (ETO, 2023)).

9.2 Open-Source Software (OSS): A Shiny Badge of Honor in Humans' Future History

"*When in doubt, one can rarely go wrong by going public."* (James E. Rogers)

When it comes to collaborative innovation, where we stand today– as Isaac Newton calls it '*standing on the shoulder of the giants*'– has not always been this tangible. 'Open-Source

[34] Greek philosopher Heraclitus is credited with this quote.

[35] CSET, based at Georgetown University, is a '*think-tank*' focused on supporting decision makers using data-driven analysis.

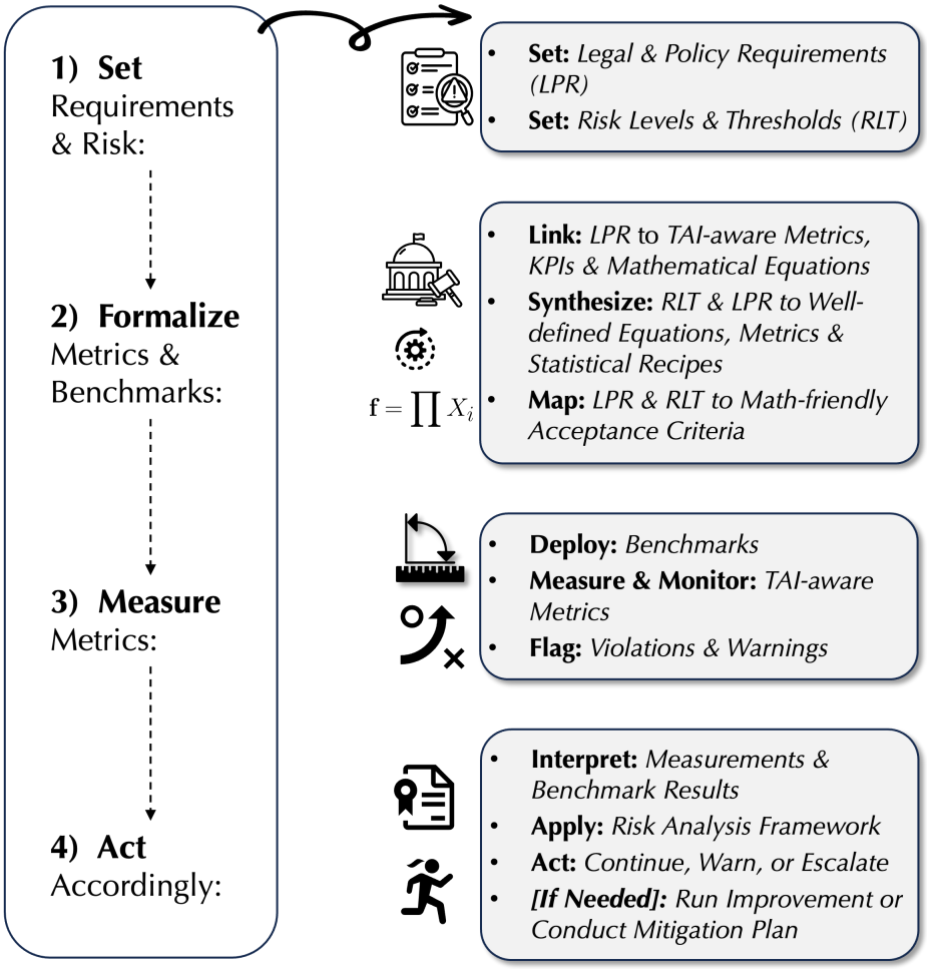

Figure 13: Our generalized framework to help enterprise and similar entities implement TAI within their organizations. For description, see discussion in § 8.4.

Table 3: Top publishing institutions in trustworthy AI research clusters identified by CSET. See (ETO, 2023) and (Toney and Probasco, 2023) for more information on how the clusters are defined.

University	Country
Arizona State University	USA
Carnegie Mellon University	USA
Massachusetts Institute of Technology	USA
University of California Los Angeles	USA
University of California Berkeley	USA
University of Notre Dame	USA
Google, LLC*	USA
Chinese Academy of Sciences	China
Nanjing University	China
Tsinghua University	China
Nanyang Technological University	Singapore
Darmstadt University of Applied Sciences	Germany
EURECOM†	France
University of Luxembourg	Luxembourg
University of Technology Sydney	Australia
University of Waterloo	Canada

* A for-profit entity.
† Graduate School and Research Center in Digital Science in Sophia, France.

Figure 14: Top scientific publishers (universities and private research labs) in 18 topics related to TAI (see text in section 9.1). Note that only one private entity is on this map, Google. Every other marker denotes a university. For definition of 18 topics in TAI and how the list of universities is compiled, we refer reader to (Toney and Probasco, 2023).

Software' (OSS) movement initiated in 1980s, currently plays a crucial and undeniable role in software and digital product life cycles. Pioneered by individual programmers, OSS ecosystem currently includes freelance developers, academia, government research facilities, for- and not-for profit companies, cf. (Korkmaz et al., 2024). The depth and breadth of OSS adoption has had major impact on entrepreneurship, innovation, and economic growth.

Table 4: A few open-source projects that had remarkable impact at a global scale. Note that the total number of contributors (the second column) for every project is extracted from its official GitHub webpage. Data extracted on February 25, 2024.)

Name	Number of Contributors*	Domain	Release Date**
Ansible	5,554	IT Automation System	2012
Bootstrap	1,387	Front-end Development	2011†
Kubernetes	3,584	Digital Product Deployment System	2014†
Linux Kernel	15,510	Operating System	1991
OpenCV	1,564	Computer Vision	2002†
OpenSSL	875	Cryptography & Network	1998
Python	2,606	Programming Language	1991
Scikit-Learn	2,815	Machine Learning Library	2007

* Reported on GitHub (as of February 26, 2024).
** Date denotes the first time that any package was shared as OSS.
† Initially launched as either a commercial or an internal-use-only software.

For example, 'Linux Operation System' (LOS) is running on all 'Top-500' supercomputers[36], and 96% of top web-servers[37]. But the best part is: Linux Kernel made it to Mars.

9.2.1 Linux Operating System 'Flying' on Mars

Ingenuity helicopter– nicknamed as Ginny– currently on **Mars** just completed its 72nd and final flight, (NASA, 2024)). Running on Linux[38] 'Operating System' (OS)– a fully open- source software– Ginny's huge success in exploring Mars is hailed by many advocates of free software systems. What is remarkable is that Ginny was built collaboratively by NASA's 'Jet Propulsion Lab' (JPL)) and utilized several open-source software during different phases. Here is what Timothy Canham– the operations lead and former software lead of the Mars helicopter project at JPL– has to say about the role of open-source in the success of Ginny:

> "*This the first time we'll be flying Linux on Mars. We're actually running on a Linux operating system. The software framework that we're using we open-sourced it a few years ago. So, you can get the software framework that's flying on the Mars helicopter, and use it on your own project. It's kind of an* ***open-source victory****, because we're flying an open-source operating system and an open-source flight software framework and flying commercial parts that you can buy off the shelf if you wanted to do this yourself someday*."(IEEE Spectrum, 2021)

[36] Based on official data as of November 2023 and released by top.

[37] Top one-million web-servers

[38] JPL used *Linaro 3.4.0*– a Linux distribution that supports Qualcomm Snapdragon processors– in Ingenuity helicopter.

Figure 15: Ingenuity helicopter (nicknamed *Ginny*) was photographed on the surface of Mars on August 2nd 2023, by another current 'resident' of Mars– the Perseverance Mars Rover. The software framework used in Ginny by NASA and JPL was based on Linux kernel– an open-source software. Credit: NASA/JPL-Caltech-ASU/MSSS.

9.2.2 Let's not Take Open-source for Granted: Hiding Scientific Discoveries for '*Job Security*' in the Past

Making it this far to this level of openness in sharing knowledge and collaborations amongst fans, enthusiasts, and academics has not been always this easy. Why? Here is a good example from history of mathematicians. '*Scipione del Ferro*', a Renaissance mathematician from Italy is credited with the first to discover an analytical solution for a subset of cubic equations of the form $x^3 + ax = b$. His ingenious solution approach led to further innovations in mathematics and complex numbers. Yet, Scipione kept this achievement secret until his deathbed, when he finally shared his notebook with his student, '*Antonio Maria Fiore*', cf. (Feldmann, 1961) and (The Editors of Encyclopedia Britannica, 2023).

This level of secrecy was common during the Renaissance era, as scholars feared competition and the potential loss of their academic positions, leading many to withhold their most **significant discoveries** as a form of "**job security**".

9.3 Open Sourcing AI: *Free-as-in-Beer vs Free-as-in-Speech*

"*When in doubt, mock the powerful, not the powerless.*" (Jon I. Lovett)

The choice of what and how 'freely' AI should be shared and available online would be shared has roots in now famous philosophy by 'GNU's Not Unix' (GNU) which articulates the term 'free software':

> "*Free software means software that respects users' freedom and community. Roughly, it means that the users have the freedom to run, copy, distribute, study, change and improve the software. Thus, "free software" is a matter of liberty, not price. To understand the concept, you should think of "free" as in "free speech", not as in "free beer". We sometimes call it "libre software", borrowing the French or Spanish word for "free" as in freedom, to show we do not mean the software is gratis.*".(GNU Project, 2024)

The movement behind 'Open-Sourcing AI' has gained momentum, mainly, thanks to existing and widely-accepted OSS ecosystem and community culture. However, inherent to the nature of AI products, existing open-sourcing software frameworks cannot capture the entire complexity associated with open-sourcing AI products. In other words, while sharing an algorithm that were utilized to build and train an AI model (along with the actual computer code in human readable form) is a good start, we argue that 'true' AI open-source requires additional components and considerations. The list below is a good example when major players consider open-source AI frameworks.

- **A Trained AI Model:** This can vary, as an AI model could be shared as an executable, binary, or byte-code, e.g. weights and biases of in a DNN.
- **(Or) Recipe to the Train AI Model**: Alternatively, an AI model's architecture, e.g. in a DNN, number of Neurons per layer, activation function type, can be shared. Yet, the actual values of weights and biases should be (re-)produced using a 'training recipe'.
- **Data:** Governance, license, meta-data, or recipes to generate (synthetic or generic) data
- **Deployment:** Includes runtime dependencies, libraries, operating systems etc. A common popular solution being the docker containers.
- **Provenance:** Providing the origin of data (training or test data) along with algorithm, recipe pipeline(s), or additions.
- **Legal:** With the highly-anticipated regulatory provisions to be enforced on AI, we argue that successful platforms such as GitHub can take the lead in educating their users on geographically-varying legal ramifications of their high-stakes AI projects. In short, licensing of any new (free) AI project may not be as easy as adding a simple **LICENSE.md** or **README.md** to the shared repository.

9.4 Where is AI Headed: A Few Insights from GitHub Trends

In the 2023 Octoverse[39] shares remarkable 2023 trends on GitHub and OSS driven by (Generative) AI GitHub Octoverse (2023):

1. **Leads:** USA, India, and Japan are leading on total number of individual contributors to Generative AI projects.
2. **India to dethrone USA:** Currently ranked as second, India is projected to dethrone USA as the largest developer community on GitHub by 2027. A Major driver has been the large-scale use of open banking system and government welfare system[40]
3. **Co-Pilots:** GitHub developers are **experimenting** and **building** their projects using AI-powered tools, e.g. code co-pilot.
4. **Paradigm shift in experimenting with AI:** AI developer and experimenters are shifting from more "traditional" libraries such as *TensorFlow* and *PyTorch* to pretrained and foundational models, LLMs, and even ChatGPT API.

[39] According to official GitHub website, 'Octoverse' is an annual report sharing the state of open-source by reporting data-driven insights and activity data collected from GitHub platform. For details on the methodology, we refer reader to Dohmke et al. (2023).

[40] Other notable examples are Mercado Libre, Latin America's largest e-commerce ecosystem, and Pix, Brazil's real-time payment infrastructure. Marcado Libre used GitHub to automate deployment and tests to aid its developers. The Central Bank of Brazil recently made Pix's communication protocols open-source.

5. **Generative AI amongst most popular projects:** For the first time, in 2023, open-source Generative AI repositories made it to the top 10 most popular projects [41]. Also, year 2023 saw the largest number of first-time contributors to OSS projects; 2.2M (120% increase from 2017).

6. **Building Cloud-ready Products**. Developers are scaling cloud-native applications using declarative languages and Git-based 'Infrastructure-as-Code' (IaC) workflows. Standardization in cloud deployments, notable surge in using Dockers and containers, and other cloud-native technologies, signaling a shift towards product-ready mentality.

7. **Open-source AI Innovation is Healthy:** Consider the top 20 open-source AI projects on GitHub in 2023: They are diverse in nature, application, and ownership (individuals or private companies). Some of most popular AI projects have been developed and maintained by individual SMEs with no affiliations to for-profit companies. This is another indicator that open-source projects in the field of AI can substantially contribute to the growth and mitigating the challenges with implementation of TAI.

10 Summary and Next Steps

We reviewed definitions of TAI (and its "synonyms") shared by entities such as UNESCO, IEEE, and NIST to consolidate most common 'features' of TAI. We suggest 'comprehending' the complex topic of TAI through the lens of characterizing its attributes or intrinsic proper- ties. In the past decade, there has been multi-disciplinary research to project concepts such as 'fairness', 'biased outcomes', 'risk and security', 'transparent', etc onto AI research. Yet, inherent complexity and subjective nature of aforementioned topics do not render a concrete '***one-size-fits-all***' TAI framework. To make matters more challenging, different countries and legal bodies have taken philosophically distinct path forward to regulate AI.

Here, we are offering a multi-prong path forward. Our main message is to first and foremost, empower the open-source movement. If history has shown us, panic is not the best guidance, and we strongly advise against over regulation which could hinder innovation and growth of the open-source community. To support our claims, we have reported exciting recent trends in AI derived from user activity data shared by GitHub in 2023. We do not underestimate the potential risks of modern AI model such as the GPT family and LLMs. Yet, to mitigate potential risks, supporting academic research and enabling open-source communities to access to AI models and compute platforms are very critical factors.

We demonstrate how existing frameworks such as the Rumsfeld Risk Matrix (RMM) can be applied to enable AI engineers plan for risks associated with the behavior of their AI systems. As EU-AI-Act has approached AI regulation through 'risk framework', it is imperative to combine existing ERM and any AI-system and its 'uncertainty level'. Having proper mapping between risk level (based on of the type of uncertainty type and its severeness) would be crucial for companies to avoid hefty fines provisioned in EU-AI-Act.

We proceed to introduce our (meta-)framework, '**Set→Formalize→Measure→Act**' to adopt and implement TAI. This is an example framework designed to enable various personas and decision makers involved in AI-product life-cycle. By nature, we aimed to have this (meta) framework generic enough so it can serve different for- or non-profit entities at various AI-product adoption level.

We hope this series can trigger enthusiasm in SMEs across different domains. We firmly believe that building efficient and helpful TAI frameworks requires an open and collaborative

[41] Ranked by 'contributor count'.

task. Areas such as TAI in judiciary systems, education, national or international security, or healthcare are only a few examples. The scale and complexities within these domains demand honest and multi-disciplinary collaboration. In part two of this series, we aim to provide more technical, statistical, and algorithmic details around TAI framework with focus on identifying proper metrics.

11 About the Authors

Mohamad M. Nasr-Azadani
Founder & Chief Executive Officer
mohamad@ImpartialZ.com
Impartial GradientZ

About: Mohamad is the founder of Impartial GradientZ, an advisory firm helping clients with bridging the gap between cutting- edge R&D in AI and Machine Learning and complex industry challenges. In the past, Mohamad served as a Principal Data Scientist at Accenture leading applied R&D in AI– ML-Ops, Differentiated Hardware and AI Model Deployment, Causal Inference, Physics-informed AI, and Digital Twins. He holds numerous patents, peer-reviewed publications, and presented at international conferences. Prior to Accenture, Mohamad received his PhD in Mechanical Engineering from University of California–Santa Barbara. His research was focused on computational fluid dynamics. He enjoyed building mathematical models of 'underwater avalanches' interacting with ocean floor topographies, running on supercomputers. Mohamad loves to learn new topics, and build AI models from scratch.

Jean-Luc Chatelain
Founder & Senior Managing Director
jlc@veraxcap.com
Verax Capital Advisors

About: Jean-Luc is the founder of Verax Capital Advisors, offering strategic guidance in AI and digital transformation. In the past, Jean-Luc served as the Global CTO for Applied Intelligence at Accenture, leading AI initiatives with a global team of 25,000 focused on AI. Jean-Luc also acted as a trusted advisor for Accenture's global 2000 clients, guiding their AI and analytics modernization journeys, and held senior executive roles at DDN and Hewlett Packard. He holds engineering and technology degrees from institutions in France, and throughout his career has created and co-invented patented technologies related to data protection, code certification, instrumentation planning, and artificial intelligence. Currently, Jean-Luc serves on advisory boards such SambaNova.ai, Writer.ai, Hyperscience, LigaData and venture capital firm Forte Ventures. He is a member of the board of directors at K2view.

A Appendix

A.1 Nomenclature

4IR Fourth Industrial Revolution

A/IS Autonomous & Intelligent Systems

AGI Artificial General Intelligence

AI Artificial Intelligence

AR Augmented Reality

CIO Chief Information Officer

COMPAS Correctional Offender Management Profiling for Alternative Sanctions

CSET Center for Security and Emerging Technology

DARPA Defense Advanced Research Projects Agency

DNN Deep Neural Network

DX Digital transformation

EAD Ethically Aligned Design

ERM Enterprise Risk Management

EU European Union

FTC Federal Trade Commission

FVVT Fairness Verification and Validation Testing

GAI Generative Artificial Intelligence

GNU GNU's Not Unix

GPAI General-purpose Artificial Intelligence

GPT Generative Pre-trained Transformer

IaC Infrastructure-as-Code

IEEE Institute of Electrical and Electronics Engineers

IoT Internet of Things

JPL Jet Propulsion Lab

LLM Large Language Model

LOS Linux Operation System

LPR Legal & Policy Requirements

ML-Ops Machine Learning Operations

ML Machine Learning

NIST National Institute of Standards and Technology

OECD Organization for Economic Co-operation and Development

OSS Open-Source Software

OS Operating System

RCA Root Cause Analysis

RLT Risk Levels and Thresholds

RMF Risk Management Framework

RRM Rumsfeld Risk Matrix

SME Subject Matter Expert

TAI Trustworthy Artificial Intelligence

UAE United Arab Emirates

UNESCO United Nation's Educational, Scientific and Cultural Organization

UQ Uncertainty Quantification

XAI eXplainable Artificial Intelligence

A.2 Guiding Principles for Trustworthy AI Released by Various Entities

A.2.1 NIST: Characteristics of a Trustworthy AI System

An agency of the United States Department of Commerce, NIST's mission is to enable innovation and competitiveness in the American industries, cf. Official Website of the National Institute of Standards and Technology. In its first published draft, NIST approaches realization of TAI through a 'Risk Management Framework' (RMF), i.e. 'AI-RMF 1.0'. The following characteristics constitute the foundation for an AI system to be considered '*trustworthy*'(National Institute of Standards and Technology (2022)):

1. Valid and Reliable
2. Safe
3. Secure and Resilient
4. Accountable and Transparent
5. Explainable and Interpretable
6. Privacy-enhanced
7. Fair– with Harmful Bias Managed

A.2.2 UNESCO: Ten Principles to Achieve Ethical AI

UNESCO held its 41st session in November 2021, Paris. With more than 190 members, UNESCO laid out **ten principals** to guide countries and private entities for the development of 'Ethical AI' with a focus on human-rights centered approach (see UNESCO (2021)):

1. Proportionality and '*Do No Harm*'
2. Safety and Security
3. Right to Privacy and Data Protection
4. Multi-stakeholder and Adaptive Governance & Collaboration
5. Responsibility and Accountability
6. Transparency and Explainability
7. Human Oversight and Determination
8. Sustainability
9. Awareness & Literacy
10. Fairness and Non-discrimination

A.2.3 IEEE: 'Ethically Aligned Design' of Autonomous & Intelligent Systems

IEEE[42]– one of the most prominent global societies of engineers and technical professionals– in 2016 released 'Ethically Aligned Design' (EAD) principles (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017)) to recommend standards in areas pertaining to 'Autonomous & Intelligent Systems' (A/IS). There has been a second version (EADv2) according to IEEE's announcement:

> "*The most comprehensive, crowd-sourced global treatise regarding the ethics of Autonomous and Intelligent Systems available today, EADv2 provides an open platform for thought leadership and action to prioritize value-driven, ethically-aligned design for autonomous and intelligent systems.*" IEEE (Accessed on 29th January 2024a)

In its mission, IEEE shares the following general principles aimed to realize 'ethical' A/IS systems:

Principle 1: Human Rights

Principle 2: Prioritizing Well-being (of humans)

Principle 3: Accountability

Principle 4: Transparency

Principle 5: A/IS Technology Misuse and Awareness of it

What is imperative about the above principles is that the 'IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems' is motivated to address these concerns pragmatically via a '*solutions-by-design*' approach[43].

A.2.4 OECD: AI Principles and Recommendations for Policy Makers

OECD is an international organization with 38 member countries. Primary mission of OECD is to stimulate economic growth and improve world trade. In doing so, OECD commits to honor democracy and market economy.

> "The OECD AI Principles promote use of AI that is **innovative** and **trustworthy** and that respects **human rights** and **democratic values**. Adopted in May 2019, they set standards for AI that are practical and flexible enough to **stand the test of time**." OECD (2023)

A.3 Example Product Requirement Document: To Build and Deploy a Trustworthy AI System for Credit Risk Score Assessment

Product Requirements Document: Creditworthiness Risk Assessment

❒ **Introduction:**

(a) **Purpose:** Estimate 'creditworthiness risk score' of individuals using AI

[42] IEEE is the world's largest technical & non-profit association of professionals and engineers whose mission has been "*dedicated to advancing technology for the benefit of humanity*" IEEE (Accessed on 29th January 2024b)

[43] Another common approach is 'solutions-by-review' which attempts to monitor and fix the ethical issues after the product is created, i.e. *post hoc* enforcement.

(b) **Scope:** Available to our clients in 42 countries in North America, Europe, and Africa

❒ **Primary Objectives:**

(a) Improve loan decision accuracy

(b) Ensure geographically-dependent legal compliance

(c) Reduce loan application processing time

❒ **Stakeholders:**

(a) Loan originators

(b) Legal and compliance team

(c) IT development team

(d) Data Science team

(e) *[External]* Loan applicants

❒ **Product Features:**

(a) Credit Risk Score Estimator:

- **Inputs:** Variable. Potential features: income, debt-to-income ratio, credit history, employment, assets, education level, demographics (only in countries where legally permitted) . . .
- **Output:** A numerical creditworthiness risk score with clear thresholds defining: "*approval*", "*rejection*", and "*needs review*" ranges.
- **AI Model:** Recommended model types are: a) Logistic Regression, b) Decision Trees, or c) Gaussian Naive Bayes. *Note: Model choice impacts explain- ability capacity. Consult legal team for further information.*

(b) Regulatory Compliance Module:

- Geo-Location identification: How applicant's location (country, potentially state/region) was utilized– if any?
- Permissible data features. Note: These features are country and/or regional dependent.
- Bias and fairness monitoring procedures. Note: Relevant metrics chosen to quantify fairness/bias and formulated recipes must be reported clearly. Exception: To comply with EU-AI-Act v0.1-2023, use EU-Fairness-Wizard internal tool.
- Model decision explainability features:
  - **Global Explanations:** Feature importance; Parity statistics; Human-comprehensible factors influencing the risk score; Uncertainty estimates
  - **Local Explanations (where required):** Provides applicant-specific reasons for their score/decision.

❒ **Reporting & Monitoring:**

- **Legal Compliance Dashboard:** Tracks key metrics across sensitive legal mandates, e.g. fairness test results across demographics as required by local laws.
- **AI Model Performance Tracking:** Monitors accuracy, drift, and any disparate outcomes.

- **Audit Logs:** Tracks all model usage, including inputs, outputs, and any regulatory disclosures.

❒ **Technical Considerations:**

- **Integration:** Interfaces with existing loan application products.
- **Security:** Adheres to the company's strict data privacy and security protocols.
- **Scalability:** Accommodates the expected growth.

❒ **Open Issues & Constraints:**

- **Legal Review:** Continuous legal counsel is needed to keep regulatory rule sets updated. Periodic external audit of AI model is highly recommended.
- **Data Availability:** Sourcing reliable data in some jurisdictions may be a challenge.
- **Explainability vs. Model Performance:** Creating explainable models might involve trade-offs with potential accuracy.

References

The list of top500 supercomputers. https://www.top500.org/lists/top500/2023/11/Accessed: February 25, 2024.

Civil Rights Act of 1964, 1964. Pub. L. No. 88-352, 78 Stat. 241 (1964).

ACM News. Japan goes all in: Copyright doesn't apply to ai training. https://cacm.acm.org/news/273479-japan-goes-all-in-copyright-doesnt-apply-to-ai-training/fulltext. Accessed on February 6, 2024.

Ashmore, Radu Calinescu, and Colin Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5):1–39, 2021.

Amy Aukema, Kate Berman, Travis Gaydos, Ted Sienknecht, Chou-Lin Chen, Chris Wiacek, Tim Czapp, and Schuyler St Lawrence. Real-world effectiveness of model year 2015–2020 advanced driver assistance systems. In *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, number 23-0170, 2023.

Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. A systematic literature review of user trust in ai-enabled systems: An hci perspective. *International Journal of Human–Computer Interaction*, pages 1–16, 2022.

Solon Barocas and Andrew D Selbst. Big Data's Disparate Impact. *California law review*, pages 671–732, 2016.

James Barrat. *Our final invention: Artificial intelligence and the end of the human era*. Hachette UK, 2023.

BBC News. Home Office drops 'racist' algorithm from visa decisions. https://www.bbc.com/news/technology-53650758, August 2020. Accessed on January 30, 2024.

David Beer. The social power of algorithms. In *The Social Power of Algorithms*, pages 1–13. Routledge, 2019.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

President Joe Biden. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House, 2023. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

J Mark Bishop. Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11:2603, 2021.

Douglas Blackiston, Emma Lederer, Sam Kriegman, Simon Garnier, Joshua Bongard, and Michael Levin. A cellular platform for the development of synthetic living machines. *Science Robotics*, 6(52):eabf1571, 2021.

Bloomberg. OpenAI's Altman sees UAE as world's AI regulatory testing ground. https://www.bloomberg.com/news/articles/2024-02-13/openai-s-altman-says-uae-could-be-an-ai-sandbox-for-the-world, February 2024. Accessed on 22nd February 2024.

Philip Bromiley, Michael McShane, Anil Nair, and Elzotbek Rustambekov. Enterprise risk management: Review, critique, and research directions. *Long range planning*, 48(4):265–276, 2015.

Benjamin S Bucknall and Shiri Dori-Hacohen. Current and near-term ai as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 119–129, 2022.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Giliberto Capano, Jeremy Rayner, and Anthony R Zito. Governance from the bottom up: Complexity and divergence in comparative perspective. *Public Administration*, 90(1):56–73, 2012.

Joseph Carlsmith. Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*, 2022.

Keith Jin Deng Chan, Gleb Papyshev, and Masaru Yarime. Balancing the tradeoff between regulation and innovation for artificial intelligence: An analysis of top-down command and control and bottom-up self-regulatory approaches. *Available at SSRN*, 2022. doi: 10.2139/ssrn.4223016.

China Law Translate. Interim measures for the management of generative artificial intelligence services (translation). https://www.chinalawtranslate.co/en/generative-ai-interim/, 2023. Accessed on 2024-02-07.

Jin-Hee Cho, Kevin Chan, and Sibel Adali. A survey on trust modeling. *ACM Computing Surveys (CSUR)*, 48(2):1–40, 2015.

Jessica B Cicchino. Real-world effects of general motors forward collision alert and front automatic braking systems. *Arlington, VA: Insurance Institute for Highway Safety*, 2018.

Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. A U-turn on Double Descent: Rethinking Parameter Counting in Statistical Learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Cristina De Fuentes and Rubén Porcuna. Predicting audit failure: Evidence from auditing enforcement releases. *Spanish Journal of Finance and Accounting/Revista Española de Financiación y Contabilidad*, 48(3):274–305, 2019.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Thomas Dohmke, Marco Iansiti, and Greg Richards. Sea change in software development: Economic and productivity analysis of the ai-powered developer lifecycle. *arXiv preprint arXiv:2306.15033*, 2023.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

ETO. Exploring trustworthy AI research with the Map of Science, part 3: Leading research institutions. https://eto.tech/blog/exploring-trustworthy-ai-research-part-three/, 2023. Accessed: February 9, 2024.

European Parliament Press. Artificial intelligence act: Deal on comprehensive rules for trustworthy ai. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai, December 2023. Accessed on March 3, 2024.

Frederik Federspiel, Ruth Mitchell, Asha Asokan, Carlos Umana, and David McCoy. Threats by artificial intelligence to human health and human existence. *BMJ global health*, 8(5), 2023.

Richard W Feldmann. The cardano-tartaglia dispute. *The Mathematics Teacher*, 54(3):160–163, 1961.

Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1):2053951719860542, 2019.

Brian Fildes, Michael Keall, Niels Bos, Anders Lie, Yves Page, Claus Pastor, Lucia Pennisi, Matteo Rizzi, Pete Thomas, and Claes Tingvall. Effectiveness of low speed autonomous emergency braking in real-world rear-end crashes. *Accident Analysis & Prevention*, 81: 24–29, 2015.

Ragnar Fjelland. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7(1):1–9, 2020.

Jane E Fountain. The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly*, 39(2):101645, 2022.

Foxglove. Home Office says it will abandon its racist visa algorithm after we sued them. https://www.foxglove.org.uk/2020/08/04/home-office-says-it-will-abandon-its-racist-visa-algorithm-after-we-sued-them/, 2020. Accessed on December 18, 2023.

Ori Freiman. Making sense of the conceptual nonsense 'trustworthy ai'. *AI and Ethics*, 3(4): 1351–1360, 2023.

Aaron French, Jung P Shim, Marten Risius, Kai R Larsen, and Hemant Jain. The 4th industrial revolution powered by the integration of ai, blockchain, and 5g. *Communications of the Association for Information Systems*, 49(1):6, 2021.

Brett Frischmann and Evan Selinger. *Re-engineering humanity*. Cambridge University Press, 2018.

CM. Froehle. The Evolution of an Accidental Meme. https://medium.com/@CRA1G/the-evolution-of-an-accidental-meme-ddc4e139e0e4, 2016. Accessed on December 18, 2023.

Vassilis Galanos. Exploring expanding expertise: artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. *Technology Analysis & Strategic Management*, 31(4):421–432, 2019.

Diego Gambetta et al. Can we trust trust. *Trust: Making and breaking cooperative relations*, 13(2000):213–237, 2000.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

GitHub Octoverse. The state of open-source and ai. https://github.blog/2023-11-08-the-state-of-open-source-and-ai/, November 2023. Accessed on 2024-02-25.

GNU Project. What is Free Software? https://www.gnu.org/philosophy/free-sw.html, 2024. Accessed: February 9, 2024.

Government of the United Kingdom. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023, 2023. Accessed on January 30, 2024.

Mark Granovetter. Economic action and social structure: The problem of embeddedness. In *The sociology of economic life*, pages 22–45. Routledge, 2018.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023.

Francesco Gualdi and Antonio Cordella. Artificial intelligence and decision-making: The question of accountability. 2021.

David Gunning. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2):1, 2017.

Thilo Hagendorff. Blind spots in ai ethics. *AI and Ethics*, 2(4):851–867, 2022.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Holistic AI. How is Brazil Leading South America's AI Legislation Efforts? https://www.holisticai.com/blog/brazil-ai-legislation-proposals, November 2023. Accessed: February 9, 2024.

George C Homsy, Zhilin Liu, and Mildred E Warner. Multilevel governance: Framing the integration of top-down and bottom-up policymaking. *International Journal of Public Administration*, 42(7):572–582, 2019.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 49–58, 2019.

IEEE. Ieee releases ethically aligned design, version 2 to show "ethics in action" for the development of autonomous and intelligent systems (a/is). https://standards.ieee.org/news/ead_v2/, Accessed on 29th January 2024a.

IEEE. Ieee vision & mission. https://www.ieee.org/about/vision-mission.html, Accessed on 29th January 2024b.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. IEEE, 2017. URL http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

IEEE Spectrum. How NASA Designed a Helicopter That Could Fly Autonomously on Mars. https://spectrum.ieee.org/nasa-designed-perseverance-helicopter-rover-fly-autonomously-mars, 2021. Accessed: February 9, 2024.

Félix Ingrand and Malik Ghallab. Deliberation for autonomous robots: A survey. *Artificial Intelligence*, 247:10–44, 2017.

C Isidore. Machines are driving wall street's wild ride, not humans. *CNN Business, CNN, Feb*, 6, 2018.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

Harvey S James Jr. The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness. *Journal of Economic Behavior & Organization*, 47(3):291–307, 2002.

Wei Jiang, Bin Han, Mohammad Asif Habibi, and Hans Dieter Schotten. The road towards 6g: A comprehensive survey. *IEEE Open Journal of the Communications Society*, 2:334–366, 2021.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

Katie Collins–CNET. UK Gov agrees to redesign 'racist' algorithm that decides visa applications. https://www.cnet.com/tech/services-and-software/uk-gov-agrees-to-redesign-racist-algorithm-that-decides-visa-applications/, 2020. Accessed on December 18, 2023.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295:103458, 2021.

Nima Kordzadeh and Maryam Ghasemaghaei. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409, 2022.

Gizem Korkmaz, J Bayoán Santiago Calderón, Brandon L Kramer, Ledia Guci, and Carol A Robbins. From github to gdp: A framework for measuring open-source software innovation. *Research Policy*, 53(3):104954, 2024.

Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access*, 2023.

Sam Kriegman, Douglas Blackiston, Michael Levin, and Josh Bongard. Kinematic self-replication in reconfigurable organisms. *Proceedings of the National Academy of Sciences*, 118(49):e2112672118, 2021.

Olya Kudina and Peter-Paul Verbeek. Ethics from within: Google glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values*, 44(2): 291–314, 2019.

Lars Kunze, Nick Hawes, Tom Duckett, Marc Hanheide, and Tomáš Krajník. Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, 3(4):4023–4030, 2018.

Nancy K Lankton, D Harrison McKnight, and John Tripp. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10): 1, 2015.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the European union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

Yann LeCun. Meta AI Chief Yann LeCun Skeptical About AGI, Quantum Computing. https://www.cnbc.com/2023/12/03/meta-ai-chief-yann-lecun-skeptical-about-agi-quantum-computing.html, 2023. Accessed on December 4, 2023.

David Levi-Faur. *The Oxford handbook of governance*. Oxford University Press, 2012.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

Albert J Menkveld. The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics*, 8:1–24, 2016.

Christian Meske and Enrico Bunde. Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 54–69. Springer, 2020.

Alok Mishra and Ziadoon Otaiwi. Devops and software quality: A systematic mapping. *Computer Science Review*, 38:100308, 2020.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3 (2):2053951716679679, 2016.

Rahul Mohanani, Iflaah Salman, Burak Turhan, Pilar Rodr´ıguez, and Paul Ralph. Cognitive biases in software engineering: a systematic mapping study. *IEEE Transactions on Software Engineering*, 46(12):1318–1339, 2018.

NASA. NASA Mars Helicopter. https://mars.nasa.gov/technology/helicopter/overview/#Five-Things, 2024. Accessed: February 9, 2024.

National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1, 2022. DOI: 10.6028/NIST.AI.100-1.

Jon Michael Raasch-Fox News. AI gives birth to AI: Scientists say machine intelligence now capable of replicating without humans, December 2023. URL https://www.foxnews.com/science/ai-gives-birth-ai-scientists-say-machine-intelligence-capable-replicating-without-humans. Accessed on January 5, 2024.

Peter A Noseworthy, Zachi I Attia, LaPrincess C Brewer, Sharonne N Hayes, Xiaoxi Yao, Suraj Kapa, Paul A Friedman, and Francisco Lopez-Jimenez. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. *Circulation: Arrhythmia and Electrophysiology*, 13(3):e007988, 2020.

OECD. The state of implementation of the OECD AI principles four years on. (3), 2023. doi: https://doi.org/https://doi.org/10.1787/835641c9-en

Official Website of the National Institute of Standards and Technology. About NIST. https://www.nist.gov/about-nist. Accessed on January 22, 2024.

Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. I lead, you help but only with enough details: Understanding user experience of co creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

Brad C Pardue. *Printing, power, and piety: appeals to the public during the early years of the English Reformation*, volume 162. Brill, 2012.

Scott Patterson. *Dark pools: The rise of the machine traders and the rigging of the US stock market*. Currency, 2013.

Thomas Philbeck and Nicholas Davis. The fourth industrial revolution. *Journal of International Affairs*, 72(1):17–22, 2018.

Shilin Qiu, Qihe Liu, Shijie Zhou, and Chunjiang Wu. Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*, 9(5):909, 2019.

Parijat Rai, Saumil Sood, Vijay K Madisetti, and Arshdeep Bahga. Guardian: A multi-tiered defense architecture for thwarting prompt injection attacks on llms. *Journal of Software Engineering and Applications*, 17(1):43–68, 2024.

Reuters. Governments race to regulate AI tools. https://www.reuters.com/technology/governments-race-regulate-ai-tools-2023-10-13/, October 2023. Accessed on 14th February 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

Randy Robertson. *Censorship and Conflict in Seventeenth-Century England: The Subtle Art of Division*. Penn State Press, 2015.

Julian B Rotter. Interpersonal trust, trustworthiness, and gullibility. *American psychologist*, 35(1):1, 1980.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1):1, 2020.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

Johannes Schneider, Christian Meske, and Pauline Kuss. Foundation models: A new paradigm for artificial intelligence. *Business & Information Systems Engineering*, pages 1–11, 2024.

Mojtaba Shahin, Muhammad Ali Babar, and Liming Zhu. Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE access*, 5:3909–3943, 2017.

Matt Sheehan. China's ai regulations and how they get made. *Carnegie Endowment for International Piece. Accessed August* 2023.

Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2459–2468, 2019.

Charlotte Stix. Artificial intelligence by any other name: a brief history of the conceptualization of "trustworthy artificial intelligence". *Discover Artificial Intelligence*, 2(1):26, 2022.

Supreme Court of the State of Arizona. Varela v. FCA US LLC, ET AL., 2022. URL https://www.azcourts.gov/Portals/0/OpinionFiles/Supreme/2022/CV200157PR.pdf. No. CV-20-0157-PR.

Georgios Symeonidis, Evangelos Nerantzis, Apostolos Kazakis, and George A Papakostas. Mlops-definitions, tools and challenges. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0453–0460. IEEE, 2022.

Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42, 2019.

Telefonaktiebolaget LM Ericsson. Follow the journey to 6G. https://www.ericsson.com/en/6g. Accessed on March 2, 2024.

Matteo Testi, Matteo Ballabio, Emanuele Frontoni, Giulio Iannello, Sara Moccia, Paolo Soda, and Gennaro Vessio. Mlops: A taxonomy and a methodology. *IEEE Access*, 10:63606–63618, 2022.

The Editors of Encyclopaedia Britannica. Scipione ferro, 2023. URL https://www.britannica.com/biography/Scipione-Ferro. Accessed 9 December 2023.

Scott Thiebes, Sebastian Lins, and Ali Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31:447–464, 2021.

Christopher Thomas and Antonio Ponton-Nunez. Automating judicial discretion: How algorithmic risk assessments in pretrial adjudications violate equal protection rights on the basis of race. *Law & Ineq.*, 40:371, 2022.

Autumn Toney and Emelia Probasco. Who cares about trust? clusters of research on trustworthy ai. Technical report, Center for Security and Emerging Technology, 2023. URL https://doi.org/10.51593/20230014b.

Pier Domenico Tortola. Clarifying multilevel governance. *European Journal of Political Research*, 56(2):234–250, 2017.

Jon Truby, Rafael Dean Brown, Imad Antoine Ibrahim, and Oriol Caudevilla Parellada. A sandbox approach to regulating high-risk artificial intelligence applications. *European Journal of Risk Regulation*, 13(2):270–294, 2022.

UK Government. A Pro-innovation Approach to AI Regulation. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#executive-summary. Accessed on February 6, 2024.

UN Human Rights Council. UN Resolution- HRC/RES/47/16 the promotion, protection, and enjoyment of human rights on the internet: resolution, July 2021. URL https://undocs.org/A_HRC_RES_47_16-EN. Accessed on: March 8, 2024.

UNESCO. Recommendation on the ethics of artificial intelligence, 2021. URL https://unesdoc.unesco.org/ark:/48223/pf0000380455.locale=en. Accessed on 2024-01-29.

USA Today. Equifax had patch 2 months before hack and didn't install it, security group says. https://www.usatoday.com/story/money/2017/09/14/equifax-identity-theft-hackers-apache-struts/665100001/. Accessed on: March 5, 2024.

Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 ieee conference on visual analytics science and technology (vast)*, pages 104–115. IEEE, 2017.

Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

Fei Yang and Yu Yao. A new regulatory framework for algorithm-powered recommendation services in china. *Nature Machine Intelligence*, 4(10):802–803, 2022.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

Harin Yoon and Soojin Jun. Ethical awareness of uxers in the loop: Ethical issues in the uxer-ai collaboration process from a ux perspective. In *Proceedings of the 25th International Conference on Mobile Human-Computer Interaction*, pages 1–6, 2023.

Yao Zhang and Qiang Ni. Recent advances in quantum machine learning. *Quantum Engineering*, 2(1):e34, 2020.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.

