Governing AI Safety through Independent Audits

Gregory Falco
Department of Civil and Systems Engineering and
Institute for Assured Autonomy,
Johns Hopkins University
falco@stanford.edu

Ben Shneiderman
Department of Computer Science and
Institute for Advanced Computer Studies
University of Maryland
ben@cs.umd.edu

Julia Badger
NASA Johnson Space Center
julia.m.badger@nasa.gov

Ryan Carrier
For Humanity
ryan@forhumanity.center

Anton Dahbura
Institute for Assured Autonomy
Johns Hopkins University
antondahbura@jhu.edu

David Danks
Departments of Philosophy and Psychology
Carnegie Mellon University
ddanks@cmu.edu

Martin Eling
Institute of Insurance Economics
School of Finance
University of St. Gallen
martin.eling@unisg.ch

Alwyn Goodloe
NASA Langley Research Center
a.goodloe@nasa.gov

Jerry Gupta
Swiss RE
jerry_gupta@swissre.com

Christopher Hart
Former Chairman
National Transportation Safety Board
chris@hartsolutionsllc.com

Marina Jirotka
Department of Computer Science
University of Oxford
marina.jirotka@cs.ox.ac.uk

Henric Johnson
Wallenberg Research Link
Stanford University
henricj@stanford.edu

Cara LaPointe
Institute for Assured Autonomy and
Applied Physics Laboratory
Johns Hopkins University
cara.lapointe@jhuapl.edu

Ashley J. Llorens
Applied Physics Laboratory
Johns Hopkins University

Alan K. Mackworth
Department of Computer Science
University of British Columbia
mack@cs.ubc.ca

Carsten Maple
WMG
University of Warwick
cm@warwick.ac.uk

Sigurður Emil Pálsson
Faculty of Physical Sciences
University of Iceland
sep@hi.is

Frank Pasquale
Brooklyn Law School
frank.pasquale@brooklaw.edu

Alan Winfield
Bristol Robotics Lab
UWE Bristol
alan.winfield@brl.ac.uk

Zee Kin Yeong
Data Innovation and Protection Group
Infocomm Media Development Authority of Singapore
yeong_zee_kin@pdpc.gov.sg

***Abstract*—Highly automated systems are becoming omnipresent. They range in function from self-driving vehicles to advanced medical diagnostics and afford many benefits. However, there are assurance challenges that have become increasingly visible in high-profile crashes and incidents. Governance of such systems is critical to garner widespread public trust. Governance principles have been previously proposed offering aspirational guidance to automated system developers; however their implementation is often impractical given the excessive costs and processes required to enact and then enforce the principles. This paper, authored by an international and multidisciplinary team across government organizations, industry and academia proposes a mechanism to drive widespread assurance of highly automated systems: independent audit. As proposed, independent audit of AI systems would embody three "AAA" governance principles of prospective risk Assessments, operation Audit trails and system Adherence to jurisdictional requirements. Independent audit of AI systems serves as a pragmatic approach to an**

otherwise burdensome and unenforceable assurance challenge.

***Index Terms*—Automated Systems, Human Control, Safety, Assurance, Governance, Design, Human-Centered Artificial Intelligence, Responsibility, Risk, Ethics**

I. Introduction

Highly automated systems (often called autonomous systems) enabled by artificial intelligence (AI) are widely used in modern society. These systems interact with people and each other in complex ways with varying degrees of human control. Such systems can add value in supporting critical infrastructure sectors such as transportation, healthcare, and financial services, but they can introduce safety risks to individuals and communities that must be addressed, particularly as they are deployed at scale. While there are many types of highly automated systems that engage AI, for purposes of this paper, the authors are concerned with those that analyze data, make decisions by engaging an algorithm and then automatically act on these decisions - ultimately having consequential impacts on society.

Increasingly visible safety incidents involving highly automated systems, such as the Airbus A330 [22] crash or two Boeing 737 MAX crashes [30] are becoming national headlines, bringing questions of highly automated system governance to the forefront of the public eye. Tesla's highly automated vehicle related safety concerns and crashes have prompted the National Transportation Safety Board (NTSB) to request that the National Highway Traffic Safety Administration to outline stricter guidelines for autonomous vehicle development and testing on public roads [29]. Such incidents are not new, but given their attention and mounting public concern, it is critical that stakeholders have assurances about such systems' performance, especially when their impact is consequential and can result in safety issues.

Government authorities and market drivers have critical roles in governing automated system assurance - especially on government-designated critical infrastructure sectors. Requirements set by these entities will be applicable to a range of systems, but will only have impact if pragmatic. A challenge is that enforcing assurance requirements for every organization is a daunting task by any singular organization. This paper outlines a regulatory mechanism to achieve assurance at scale: the Independent Audit of AI Systems (IAAIS - pronounced "eyes"). The proposed audit framework could embody the authors' proposed "AAA" governance principles:

1) Prospective **A**ssessments before highly automated systems are implemented
2) **A**udit trail to analyze failures and help assess accountability
3) System **A**dherence to jurisdictional requirements.

An independent audit framework, centered around the AAA governance principles, intends to help preempt, track, and manage safety risk while encouraging public trust in highly automated systems. Such an audit would furnish managers, manufacturers, lawmakers, and insurers with operational data, expectations, and an operational baseline for highly automated systems so they can enable human responsibility and control. Since enforceable principles must capture a range of use cases and risk considerations, the authors represent interdisciplinary fields of study and practice, including computer science, systems engineering, human-computer interaction, law, business, public policy, and ethics. The focus of the authors' argument is on consequential, life-critical safety applications that are widely used by major organizations, such as those in healthcare, transportation, and financial trading. Because of the differences across application domains, each industry sector will have to determine for which AI systems the AAA governance principles are necessary and tailor the principles to fit the safety concerns of their sector. For purposes of this discussion, the authors define safety in the context of physical, emotional and financial safety as described in Maslow's Hierarchy of Needs [19]. While, the AAA principles described could potentially be useful to help govern other impacts of highly automated systems, such as those on the environment or social justice, the authors choose to focus the discussion on safety as perhaps the most immediate concern for highly automated systems.

II. Enforceable Governance

Given the vast array of use cases, highly automated systems range considerably in the extent of their intelligence and degree of human control. Consequentially, their risks also vary from controlled and isolated malfunctions to cascading multi-system failures. Human-centered artificial intelligence (HCAI) provides increasing levels of automation and human control; therefore enabling reliable, safe, and trustworthy systems [27]. By effectively requiring independent audit governmental authorities (courts, government agencies) and market drivers (insurers and audit firms) could monitor for and facilitate HCAI. Governance to address both cyber and physical risks must be the result of convergence research across disciplines, enabling broad application of the governing mechanisms across use cases [7]. Multidisciplinary HCAI-reaffirming governance principles could yield highly automated systems that gain public confidence and acceptance.

A. Governance Principles

Various industry, academic, government bodies, and not-for-profit organizations have proposed AI principles in recent years, with many sharing core concepts and mechanisms [1], [11]. A recent survey of 84 sets of ethical principles in AI [16] found these principles appeared most often: *transparency* (in 73), *justice & fairness* (in 68), and *non-maleficence* (in 60). Whilst these principles are very worthy, they do not help practically with the design, development and use of automated systems. Principles alone are insufficient to address AI risks [20], [37]; an increased focus on enforceable governance, and corresponding practices and processes, would help reduce, mitigate, and control risks [15]. Wrapping principles within an independent audit framework could streamline the adoption of highly automated governance by simplifying how market drivers and government authorities encourage regulation. The

baseline AAA principles proposed here are actionable and widely applicable; thus appropriate for a broadly deployed independent audit framework.

B. Independent Audit

Financial audit and accounting provides an interesting model to emulate for the governance of AI. When Generally Accepted Accounting Principles (GAAP) were put in place in 1973, the relevant regulatory authority—the U.S. Securities and Exchange Commission (SEC)—mandated their adoption by publicly traded companies in less than 18 months. The resulting infrastructure of trust is so robust that, with the exception of malfeasance and fraud (like Enron 2001), financial reports produced by independent auditors are trusted. IAAIS would employ a consensus-driven, transparent, and stakeholder-inclusive process to craft existing laws, guidelines, and best practices into implementable, measurable, binary (compliant/non-compliant) audit rules. The system would protect corporations' intellectual property and innovations from excess transparency while protecting society with a liability shield from the independent auditor. Internal Assessments and Audit could become annual processes embedded in internal controls for corporations, within a framework of agile and responsible governance [37]. Courts and government agencies would have the ability to institutionalize audit and expand requirements in law as needed. Under IAAIS, those same laws, adapted into the audit rules, would necessitate proactive compliance to achieve a successful audit. The internal risk controls and audit process can drive Adherence to local laws by lowering risk with affirmed compliance and by providing insurers with data for thoughtful underwriting and pricing.

C. Principle 1: Assessments

By proactively identifying and enumerating potential risks to public safety, and finding acceptable methods of mitigating those issues (analogous to product disclosure), developers and operators can build the public's confidence in highly automated systems. These assessments must address the full diversity of individuals and groups that might be users of or impacted by the system, particularly since developers and operators might be unintentionally blind to potential risks or impacts. For example, a major reason that commercial aviation is so safe is the extent to which collaboration across manufacturers, operators, employees, regulators and researchers is used to identify and address potential safety issues. Transparency of known risks, as well as steps taken to mitigate those risks, is critical even if the algorithms within the highly automated system might not be explainable or transparent. Moreover, if the highly automated system uses adaptive or learning algorithms, then the assessment serves as a mechanism to monitor and manage potential assumptions that become outdated or inaccurate as time passes for the model (model drift). Associated techniques are required to identify, document and mitigate these risks. Assessments also serve to identify design tradeoffs, which reveal the strengths and weaknesses of alternatives [10]. Formal methods for verifying safety are especially valuable for specific features that are well understood [18], [31]. However, the state space is often poorly defined for highly automated systems, which calls for other methods of conducting assessments such as stress tests [32] [8]. Relevant information about the assessment techniques will ultimately need to be included in product specification requirements to provide evidence that risks were mitigated.

Standards for risk assessment are well established in safety-critical systems: for instance ISO-13489:2015 *Safety of machinery* applies to safety-related parts of a control system and IEC 31010:2019 *Risk management – Risk assessment techniques* is a standard setting out "guidance on the selection and application of techniques for assessing risk in a wide range of situations"[1]. In robotics, BS8611-2016 *Guide to the ethical design and application of robots and robotic systems* [2] provides guidance on how designers can undertake an ethical risk assessment of their intelligent robot, and mitigate any ethical risks so identified. "At its heart is a set of 20 ethical hazards and risks, and advice on measures to mitigate the impact of each risk is given, along with suggestions on how such measures might be verified or validated" [34]. Broadly, these standards are structured along the axes of the probability and severity of a given harm. This allows construction of a risk hierarchy and a range of mitigation measures, including determining the level of human oversight.

One mechanism that addresses both the enumeration of possible risks and associated standards to mitigate these is specifications or a "building code" for highly automated systems. Similar to building codes for architectural design and structural engineering, highly automated systems require minimum specifications to be assured. Such specifications for highly automated systems could serve as a template to help developers disclose and begin addressing the risks of their systems. Examples of "building codes" as described have been proposed for generic software and for medical devices [13], [17]. However, specifications for highly automated systems have yet to be developed and is an opportunity for future research.

Recent industry efforts have focused on systematic methods for documenting key aspects of systems such as Microsoft's datasheets for datasets to describe the training data, Google's Model Cards to describe the algorithmic model, IBM's FactSheets to describe all aspects of a system, and other proposals to clarify how systems would provide explanations. These are beginning to be adopted, which will lead to rapid improvements in ways to document systems.

While not existing practice, stakeholder feedback on automated system operations is a necessary part of assessments. This could take shape as a standardized review for how the highly automated system meets users' safety expectations. Including stakeholder input into assessments could help to improve the perceived transparency of the governance process. Further, it would augment the assessment with contextual in-

[1]https://www.iso.org/standard/72140.html

formation about performance which may have been otherwise omitted from the safety assessment.

D. Principle 2: Audit Trail

A means to capture the context of failures with high fidelity data is required so that those failures can be rigorously examined and accountability can be assigned. This is a concept better known in technical circles as "traceability" or requirements tracing, which in turn facilitates transparency and explainability [3], [25]. An audit trail for highly automated system operations, for instance, can provide high-fidelity data to improve traceability, and, by extension, enable accountability. Analysts could thereby either identify risks using real-time monitoring and analysis, or provide post-event visibility into the context surrounding the accident. If data regarding highly automated system accidents, including near-misses, were stored in a transparent, publicly available data repository, they would be a valuable source for researchers, authorities, and developers.

The aviation industry has successfully established audit trails for their systems using "black box" flight data recorders (FDRs). These FDRs have played a pivotal role in making aviation remarkably safe considering the complexity of their systems and processes. The demonstrable efficacy and value of FDRs in understanding accidents [12] suggests that adopting something analogous is warranted. This has previously been proposed for robotics [36] and other highly automated systems [40] [9]. While the application and measurements taken may vary, the original intent of the "black box" remains consistent: collect evidence of systems actions and the surrounding context for analysis after near misses and failures - which must be defined specific to the use case. However, effort will be required to implement FDRs into different contexts.

Accident investigation is not simply a matter of collecting data from a black box or equivalent data logger. It is a human process that involves collecting witness testimony and forensic evidence, then – alongside data from the black box – interpreting all of the evidence to discover what happened, why it happened, and what must be done to avoid it from happening again [35]. At present, accident investigation in both social robotics and HCAI is hindered by the lack of both standard specifications for a black box, and processes for investigating accidents. Both must be in place, alongside transparent and trusted accident investigation agencies, before highly automated systems can begin to earn the confidence that aviation enjoys. FDRs will vary across industry. Requirements for self-driving cars are emerging because of the U.S. NTSB guidance about what is needed to conduct retrospective analyses of car crashes. Medical device FDRs will need to record time-stamped information about every keypress with data from sensors to capture the status of the devices. Financial trading systems have already implemented recording systems for each trade for basic business needs, but more information may be needed about how machine aided decision making was performed.

While the primary purpose of FDRs is accident investigations, "black boxes" have been proven useful in other capacities. Data collection and subsequent analysis could yield continuous improvement, although every change will require fresh verification and validation tests. There is good evidence that reporting and acting on near-misses significantly improves safety [33].

Further, FDR data from different manufacturers has been shared externally for analysis in the aggregate to help identify broader trends. Given the complexity and variety of highly automated systems, lessons learned from one system may not always provide direct help for other systems; however some degree of meta-analysis could be useful to improve future system design. These benefits for the systems can only be realized if the information is made publicly available in a suitably anonymous and responsible form. This requires a markedly different approach than previously utilized for cyber-attack disclosure, where a fragmented market of cyber-data providers presents accessibility challenges to researchers. Mandatory, responsible public disclosure of anonymous data would improve upon what is provided by the aviation industry. In aviation today, individual incident report disclosure is voluntary and only the analysis of aggregate data, compiled by the MITRE Corporation, is made public to the extent approved by the Commercial Aviation Safety Team (CAST) that collects and analyzes the data. Anonymous, publicly responsible disclosure would provide transparency and the potential for analysis by independent researchers, while furnishing increased public trust.

Since traceability is context-specific, black box designers will have to choose what data to collect, based on risks of failure and/or severity of harm in each industry. The maturity of the highly automated systems will influence each industry's regulatory requirements [25]. Independent Audit of AI Systems would heavily leverage such a black box given the high-degree of traceability and transparency it affords.

E. Principle 3: Adherence

Highly automated systems will function in varied jurisdictions, each with unique rules and operating requirements. Designers will have to comply with these requirements, including geographic (e.g., municipality) and sectoral (e.g., healthcare) boundaries. For example, an assisted living robot may be located in a specific city with one set of privacy requirements, while also subject to healthcare sector regulations due to its healthcare role. Customizing highly automated systems to suit geographic and sector-specific requirements is critical to their integration to society.

An independent oversight board developed to adjudicate over questionable adherence to rules illuminated by an independent audit could help alleviate these concerns [27]. The authors propose oversight boards per industry sector on a national level, which will help address the nuances of each and their highly automated system requirements. The development of industry guides can help provide industry best practices, which simplifies the process of standardizing and

adopting baseline principles. For instance, in Singapore, the Monetary Authority of Singapore introduced financial sector-specific guides for highly automated systems guided by the overarching and sector-agnostic Model Framework [21]. In addition, industry guides such as the "Implementation and Self-Assessment Guide for Organizations", produced jointly by the World Economic Forum in conjunction with the Info-communications Media Development Authority of Singapore, highlight industry best practices and guide organizations in assessing their adoption of responsible practices in respect of highly automated systems [39].

Given that aforementioned systems are likely to operate across borders, where relevant, sector boards may contain committees or chapters to see to unique local requirements at each geographic level. For example, a vehicle oversight board may be established on a national level, where the state of Nevada or the region of New England may opt to organize a committee/chapter and overlay its own perspective on top of the national board. A drawback of the approach described is the necessary proliferation of regulatory bodies, which requires appropriate resources to function; however it is the authors' opinion that the benefits to many smaller, potentially more agile sector regulators compared to a singular regulator (à la the SEC) would outweigh these costs over the long term. While the authors assume governance will generally be on the national level and potentially augmented locally, international coordination can still be useful as is currently being pursued through UNESCO [38]. Regional policies, such as the Global Data Protection Regulation (GDPR) and ISO standards, can equally have global impacts.

Alternatively, more aggressive measures can be taken to assure the behavior of highly automated systems. Provisions can be made on local networks to enforce requirements of a given jurisdiction [6]. Establishing enforcement measures on the network supports the attributes of jurisdictional requirements and will help to manage the highly automated systems so that that they operate safely. For example, the US, Canada, UK, Sweden, Switzerland, Iceland, and Singapore (represented by the authors) have different perspectives on safety. The network thus becomes an operating system for specific use cases and locations. Rules of a given jurisdiction will be compulsory for the highly automated systems on the network, which could help drive compliance. Audits could be oriented, in part, around the extent of adherence to such rules, as described further below. To be effective, the network must be highly resilient so that enforcement mechanisms are minimally influenced by inevitable attacks.

III. Enforcement Vehicles for Independent Audit

There are three potential enforcers of independent audit: insurers, courts and government agencies.

A. Insurers

Insurers rely on a standard set of expectations to provide insurance. Today there are no such standardized expectations of highly automated systems due to the lack of baseline requirements or specifications set by enforcers. Insurers also rely on clear "attribution risk" for a system, which involves understanding the cause and cost of an accident. Attribution becomes more complicated for highly automated systems. For example, suppose a self-driving vehicle has a flat tire which gets repaired at a local garage, but that repair throws off the sensors, resulting in an accident. Attribution requires a "chain of custody" for this self-driving system, where each "custodian" is insured and accredited to work with that highly automated system.

IAAIS could benefit insurers by providing documentation for insurers of an organization's AI baseline risk, and also providing a clear audit trail–attribution risk assessment. In turn, insurers could help to encourage the adoption of independent audit by requiring both risk Assessments and real-time, continuous Audit capabilities and logs before insuring highly automated systems - principles that would be reflected in IAAIS. Both would be pivotal to help insurers know the potential risks, determine their probability of occurring, and assess the damage, cause of, and responsibility for any accident. If insurers required IAAIS, independent audit and its embodied principles would become widely adopted because most companies rely on insurance as a risk management tool. Further, insurers can also help to drive certification programs associated with an audit (whose role in encouraging the implementation of the principles is covered below) by requiring certification for certain types of policies, such as insuring a company that operates a fleet of autonomous vehicles or deploys highly automated systems at scale in its products, services or operations.

B. Courts

Some firms will fail to engage in the Assessments necessary to avoid disastrous outcomes. Tort lawsuits will follow, with plaintiffs demanding damages for firms' failures to meet the relevant standard of care. Courts will need to develop standards adequate to the new technological environment, which could include self-regulation in the form of independent audit. As they do so, they will effectively set nuanced and contextualized standards for the deployment of AI. Imagine a self-driving car which runs over and kills a pedestrian whom its sensor systems failed to recognize. Courts may decide that the standard of care for deploying an autonomous vehicle is to keep in the front seat a "guardian driver" who can take over control when the vehicle fails to notice a pedestrian—and that deviation from this standard of care results in liability for the designer, developer, owner and/or operator of the autonomous vehicle. If the technology improves, standards of care may essentially prescribe other measures, such as ensuring that up-to-date data sets are being fed into the machine learning algorithms behind the vehicle's operation.

As courts develop such evolving standards of care, they will also face questions of legal irresponsibility. Parties will naturally contest on whom liability should fall. For example, in medicine, the doctrine of "competent human intervention" has shifted liability away from those who make devices and

toward the professionals who use them. As courts address these and other forms of legal irresponsibility, they will need to develop nuanced doctrines to clarify who is held accountable for foreseeable, preventable errors. This is especially true given a focus on HCAI, where humans will be in the loop. In cases of contest, IAAIS will be critical for purposes of transparency.

Alternatively, courts could serve in a different capacity by supporting a model similar to workers' compensation for victims who are injured when something goes wrong. Such a model does not require the need to prove fault, thereby saving time, money and enabling the faster compensation of victims. An approach like this could improve AI safety and performance by enabling the developers and users of the automation to learn from their mistakes, documented in an independent audit, rather than hiding them for fear of litigation.

C. Government Agencies

Policymakers are currently struggling to keep pace with the speed of technological development. Legislators have been hesitant to pass broad statutes, as they are fearful of inhibiting growth and innovation. However, increasingly there is public demand for policy interventions and protections regarding certain digital technologies. Some fields may never gain traction if customers cannot be assured that someone will be held accountable if a highly automated system catastrophically fails.

An early example of policymaker guidance on (but not enforcement of) AI governance was in 2018 when The European Union Agency for Cybersecurity (ENISA) described various security considerations for automated systems that need to be addressed such as detection of rogue or unauthorized systems, hijacking and misuse, interference, transparency and accountability and adherence to security principles [4]. Separately, in 2019 the European Union's High Level Expert Group published "Ethics Guidelines for Trustworthy AI" describing three conditions to be met: the system should be lawful, ethical and robust, from both a technical and a social perspective [14]. In April 2021 the European Commission followed up with a proposed legal framework for AI regulation [5]. The Personal Data Protection Commission of Singapore's "Model AI Governance Framework" (Model Framework) released in 2019 also sets out recommended practices to private organizations in implementing the ethical principles of fairness, explainability, transparency and human-centricity across the adoption lifecycle of highly automated systems [24]. In 2021, the United States' National Security Commission on Artificial Intelligence released its Final Report, discussing AI-enabled weapon system governance and proposed recommendations to enable public trust through improving transparency, developing standards and performance metrics and establishing a standing body of multidisciplinary experts to advise agencies on responsible AI use [26].

The proposed independent audit and associated governance principles could provide a baseline from which policymakers can build or extend their automated systems policy platform. Policymakers could directly write laws with defined penalties or could equally establish incentives pertaining to conducting an independent audit. Given that highly automated systems are still relatively new, it is preferable to establish incentives to encourage IAAIS as regulators continue to improve their understanding of risks in specific applications. There is much to learn from extant efforts to audit data in the national security and finance sectors; templates for good auditing methods abound [23]. The tax system can also encourage better practices. Just as legislators encouraged the development and purchase of renewable energy systems and reduced carbon emission vehicles with tax benefits to companies, similar benefits could come to companies that demonstrate that they perform independent audits. Direct public support of new technologies could include these provisions, following the example of "meaningful use" conditions on subsidies to electronic health records that were part of the 2009 U.S. stimulus package (the American Reinvestment and Recovery Act).

While the legislative process is typically slow, agencies can specify enforcement means that have more immediate effects. Regulating agencies can interpret existing statutes in order to promote IAAIS. Regulators such as Japan's Financial Services Agency, China's State Electricity Regulatory Commission, and the UK's Health and Safety Executive (HSE), already have requirements related to how certain systems are used in their respective industries. As regulated sectors make increased use of highly automated systems, adopting specific baseline requirements such as conducting an independent audit will help to keep pace with rapidly evolving technology. Considering regulators hold the power to fine and even shut down non-compliant organizations, their adoption of these principles could transform the safety posture of highly automated systems for an entire industry. Regulators would be particularly helpful in enforcing the AAA principle of Adherence, as they could help define requirements for their sector.

IV. Anticipated Benefits of IAAIS

Establishing a baseline for enforceable governance of highly automated systems through IAAIS will potentially have far-reaching impact. While the AAA framework will require testing and refinement with active highly automated systems, we anticipate that it will have the benefits of encouraging both ethics and accountability.

A. Encouraging Ethics

A core challenge for ethical use of highly automated systems is assurance that the system will perform in ways that accord with the users' values. In practice, values are rarely explicitly represented in a highly automated system, but rather are implemented through a range of design, development, and deployment choices. For example, an aircraft autopilot does not explicitly represent "save the lives of passengers," but instead implements that value through its design features and pilot controls. When deployed incorrectly, the autopilot can cause the opposite of the value it is meant to espouse. As

a result, users (and others) can often struggle to determine the values implemented in a technology, and so determine whether they should use it. Alongside emerging approaches to values-based design [15], [28], the proposed IAAIS would provide much of the information and oversight required to have assurance about the behavior of systems, and thereby enable technology to be used in a more ethical and responsible manner.

B. Encouraging Accountability

While IAAIS does not directly hold designers, developers, owners and/or operators of highly automated systems to a given ethical standard, independent audit can be a useful tool in holding organizations or individuals accountable for flagrant decisions. Equally, IAAIS auditors, while currently not legally liable for falsifying information or inaccurately portraying the safety of an organization's highly automated systems, could face considerable reputational harm when the system audited has a safety failure. As the highly automated systems audit landscape evolves, auditors could become licensed to perform services by professional societies, trade organizations or federal entities where unethical conduct, potentially resulting from the principal-agent problem, could lead to being disbarred or fines [34]. Such punitive actions would ideally outweigh any potential financial benefit offered to falsify audit claims, as is largely the case today for financial auditors. As evidenced from the subprime mortgage crisis that contributed to a global economic recession between 2007 and 2011, without accountability for corporate malfeasance, the public could be seriously harmed. Traceability and the threat of accountability if something does not work as intended could facilitate ethical decision-making with regards to designing, developing or operating highly automated systems.

V. Conclusion

Introducing the AAA governance principles via an independent audit, can foster the risk awareness, responsible development and thoughtful utilization of highly automated systems. The robust discussion of AI governance principles and the use of IAAIS will promote safe highly automated systems. While other frameworks described a comprehensive library of aspirations for AI governance, the AAA principles support an independent audit that is actionable and enforceable. Such measures are necessary to foster trust in the developers, operators and regulators of such systems. Broad adoption of IAAIS can raise the perceived transparency of these systems and provide a dataset on which regulators and enforcers can build increasingly robust regulatory requirements. If users, watchdogs, and government agencies know the risks, are able to track system operations, and have assurance that systems are operating as intended, the public will have greater confidence in using highly automated systems.

VI. Acknowledgements

Special thanks to the participants of the CCC Workshop on Assured Autonomy for their ideas, inspiration and discussion which contributed to this paper. [23]

References

[1] Brundage, M., et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213* (2020).

[2] BSI. *BS8611:2016 Robots and robotic devices, Guide to the ethical design and application of robots and robotic systems*. British Standards Institute, 2016.

[3] Cummings, M., and Britton, D. Regulating safety-critical autonomous systems: past, present, and future perspectives. In *Living with Robots*. Elsevier, 2020, pp. 119–140.

[4] Drogkaris, P., and Bourka, A. Towards a framework for policy development in cybersecurity: Security and privacy considerations in autonomous agents. Report, European Union Agency for Cybersecurity, Heraklion, December 2018.

[5] European Commission. Regulation of the european parliament and of the council laying down harmonized rules of artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Tech. rep., Brussels, April 2021.

[6] Falco, G. A smart city internet for autonomous systems. *2020 Symposium on Security and Privacy Workshops (SPW)* (2020), 215–220.

[7] Falco, G., et al. Cyber risk research impeded by disciplinary barriers. *Science 366*, 6469 (2019), 1066–1069.

[8] Falco, G., and Gilpin, L. A stress testing framework for autonomous system verification and validation (v&v). *IEEE International Conference on Autonomous Systems* (2021).

[9] Falco, G., and Siegel, J. E. A distributed "black box" audit trail design specification for connected and automated vehicle data and software assurance. *SAE International Journal of Transportation Cybersecurity and Privacy 3*, 11-03-02-0006 (2020).

[2]For correspondence relating to this manuscript, please contact Gregory Falco at falco@jhu.edu.

[3]The authors declare the following competing interests: Gregory Falco is a consultant for the World Bank Group on autonomous vehicle regulation and is a ForHumanity fellow; he thanks the U.S. National Institute for Standards and Technology (NIST), the Icelandic Fulbright Commission and the National Science Foundation for research funding. Julia Badger is a US government civil servant working for NASA where all her funding comes from NASA. Ryan Carrier is the Executive Director of ForHumanity, a registered 501(c)(3) not-for-profit organization. Anton Dahbura is a member of the Maryland Cybersecurity Council, established by the Maryland State Legislature to work with the National Institute of Standards and Technology and other federal agencies, private sector businesses, and private cybersecurity experts to improve cybersecurity in Maryland. David Danks is an external member of the Salesforce Advisory Council on Ethical Humane Use of Technology. Alwyn Goodloe is a US government civil servant working for NASA where all his funding comes from NASA; he is a voting member of SAE 34 working group on AI in Aviation that is writing guidelines (similar to DO-178C) for AI in aviation. Henric Johnson is the Swedish Science and Innovation Counsellor to the U.S. Ashley Llorens is Vice President and Global Outreach Director at Microsoft Research and currently serves as the Science Representative on the steering committee of the Global Partnership on Artificial Intelligence. Alan K. Mackworth is a Director of Minerva Intelligence Inc; he is a member of the Centre for AI Decision-making and Action (CAIDA) Steering Committee, the AI network of BC (AInBC) Board, and the The Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE) International Advisory Board. Ben A. Shneiderman is a ForHumanity fellow. Alan Winfield sits on British Standards Institute committee AMT/010/01 Ethics for Robots and Autonomous Systems, the executive committee of the IEEE Standards Association Global Initiative on Ethics of Autonomous and Intelligent Systems, and the WEF Global AI Council; he is a member of the Advisory Committee of robotics company Karakuri Ltd. Authors not mentioned have no competing interests. ZK Yeong leads the development and implementation of Singapore's AI Governance Framework, is a member of OECD's Network of Experts in AI and the Global Partnership on Artificial Intelligence's expert working group on Data Governance.

[10] Fischer, G. Design trade-offs for quality of life. *ACM Interactions XXV 1* (2018), 26–33.

[11] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, 2020-1 (2020).

[12] Grossi, D. R. Aviation recorder overview. In *International Symposium On Transportation Recorders, Arlington, Virginia* (1999).

[13] Haigh, T., and Landwehr, C. Building code for medical device software security. *IEEE Cybersecurity* (2015).

[14] High-Level Expert Group on AI. Ethics guidelines for trustworthy AI. Tech. rep., European Commission, Brussels, April 2019.

[15] IEEE. Ethically Aligned Design: A vision for prioritizing human well-being with autonomous and intelligent systems, 1st edition, 2019.

[16] Jobin, A., Ienca, M., and Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1* (2019), 389–399.

[17] Landwehr, C. E. A building code for building code: putting what we know works to work. In *Proceedings of the 29th Annual Computer Security Applications Conference* (2013), pp. 139–147.

[18] Mackworth, A. K., and Zhang, Y. A formal approach to agent design: An overview of constraint-based agents. *Constraints 8*, 3 (2003), 229–242.

[19] Maslow, A. H. A theory of human motivation. *Psychological review 50*, 4 (1943), 370.

[20] Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* (2019), 1–7.

[21] Monetary Authority of Singapore and the Fairness, Ethics, Accountability and Transparency (FEAT) Committee. Principles to promote fairness, ethics, accountability and transparency (feat) in the use of artificial intelligence and data analytics in Singapore's financial sector. Report, Monetary Authority of Singapore, Singapore, November 2018.

[22] Oliver, N., Calvard, T., and Potočnik, K. The tragic crash of flight AF447 shows the unlikely but catastrophic consequences of automation. *Harvard Business Review* (2017).

[23] Pasquale, F. *The black box society*. Harvard University Press, 2015.

[24] Personal Data Protection Commission of Singapore. Model artificial intelligence governance framework, 1st edition. Report, Infocomm Media Development Authority of Singapore, Singapore, January 2019.

[25] Personal Data Protection Commission of Singapore. Model artificial intelligence governance framework, 2nd edition. Report, Infocomm Media Development Authority of Singapore, Singapore, January 2020.

[26] Schmidt, E., et al. Final report. Tech. rep., National Security Commission on Artificial Intelligence, Arlington, Virginia, March 2021.

[27] Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction 36*, 36 (2020), 495–504.

[28] Spiekermann, S., and Winkler, T. Value-based engineering for ethics by design, 2020.

[29] Sumwalt, R. Docket no. dot-nhtsa-2020-0106. Letter, National Transportation Safety Board, Washington D.C., USA, February 2021.

[30] Sumwalt, R., Landsberg, B., and Homendy, J. Assumptions used in the safety assessment process and the effects of multiple alerts and indications on pilot performance. Report, National Transportation Safety Board, Washington D.C., USA, September 2019.

[31] Tomlin, C. J., Mitchell, I., Bayen, A. M., and Oishi, M. Computational techniques for the verification of hybrid systems. *Proceedings of the IEEE 91*, 7 (2003), 986–1001.

[32] Topcu, U., et al. Assured autonomy: Path toward living with autonomous systems we can trust. *arXiv preprint arXiv:2010.14443* (2020).

[33] Williamsen, M. Near-miss reporting A missing link in safety culture. *Professional Safety 58* (2013).

[34] Winfield, A. Ethical standards in robotics and AI. *Nature Electronics 2*, 2 (2019), 46–48.

[35] Winfield, A., et al. Robot Accident Investigation: A case study in responsible robotics. In *Software Engineering for Robotics*, A. Cavalcanti, B. Dongol, R. Hierons, J. Timmis, and J. Woodcock, Eds. Springer, Cham, 2021.

[36] Winfield, A., and Jirotka, M. The case for an ethical black box. In *Annual Conference Towards Autonomous Robotic Systems* (2017), Springer, pp. 262–273.

[37] Winfield, A., and Jirotka, M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Phil. Trans. R. Soc. A 376*, 20180085 (2018).

[38] World Commission on the Ethics of Scientific Knowledge and Technology. Preliminary study on the ethics of artificial intelligence, 2019.

[39] World Economic Forum. Companion to the model AI governance framework – implementation and self-assessment for organizations. Report, Infocomm Media Development Authority of Singapore, Singapore, January 2020.

[40] Yao, Y., and Atkins, E. The smart black box: A value-driven high-bandwidth automotive event data recorder. *IEEE Transactions on Intelligent Transportation Systems* (2020).