




Applying Human Cognition to Assured Autonomy

Mónica López-González 

Institute for Human Intelligence, Baltimore, MD, USA
monica@ihintelligence.org

Abstract. The scaled deployment of semi- and fully autonomous systems undeniably depends on assured autonomy. This reality, however, has become far more complex than expected because it necessarily demands an integrated tripartite solution not yet achieved: consensus-based standards and compliance across industry, scientific innovation within artificial intelligence R&D of explainability, and robust end-user education. In this paper I present my human-centered approach to the design, development, and deployment of autonomous systems and break down how human factors such as cognitive and behavioral insights into how we think, feel, act, plan, make decisions, and problem-solve are foundational to assuring autonomy.

Keywords: Assured autonomy · Artificial intelligence · Human factors · Trust · Explainability · Autonomous vehicles

1 Introduction

Artificial intelligence (AI)-based technologies, methods, and applications are increasingly entering all industries and domains, and their negative effect on individuals and society at large is no secret. From biased facial recognition algorithms [1] and medical AI devices [2] to the abuse of natural language processing systems [3] and resulting deaths from AI-enabled vehicles [4], to name a few, the promise of AI is not without risks and challenges. As a result, the concept of ‘assured autonomy’ –or ‘trustworthy AI’ as referred to by the European Commission– is capturing the ears of governments, international organizations, and industry alike [5–8]. While new autonomous systems are being built, trust in the new technology, however, is declining [9, 10]. Add the current COVID-19 pandemic, economic crisis, public outcry over systemic racism, and disintegration of democratic stability across the globe and trust is at a new low across a swath of sectors [11].

Notwithstanding, trust is a vital social process that helps us to cooperate with others and form a relationship [12]. Inherently risky due to the unpredictability of others’ intentions, we nonetheless learn to accept vulnerability and depend on one another to produce positive, mutual advantages [13]. The success of this dependency relies on reciprocity and the gains perceived from such. For example, you offer me something with supposed ‘x’ characteristic to improve my life, I perceive or at least hope in your honesty, and I reciprocate your offer by engaging with that something because I expect ‘x’ to be true. In the context of assured autonomy, Company A offers me an AI-enabled

autonomous system ‘a’ with ‘x, y, and z’ characteristics, I believe in the integrity of the company, and I buy system ‘a’ because I expect ‘x, y, and z’ as true. Inherent in this business transaction example is my confidence in Company A providing me with the claimed system’s performance capability. Problems arise when the system does not perform as intended, and explainability of such failure is left unanswered. The lack of explainability suggests a shortfall in Company A’s standards of ethical behavior and accountability, and my trust in the company accordingly diminishes.

In this paper, an expansion of [14], I argue that we need a human-centered approach to the design, development, and deployment of AI-enabled autonomous systems if we are to fully trust the capabilities of these systems. Integrating essential human cognitive intelligence characteristics during design and development, and identifying potential negative impacts during deployment, along with mitigation strategies prior to the system’s actual deployment in the field, can aid in setting a clear standard of ethical behavior and accountability from the onset. Using the autonomous vehicle industry as a backdrop, I delineate how assured autonomy –and the consequent regaining of the public’s trust in AI– will depend on a united and cross-disciplinary effort in (a) setting consensus-based standards and compliance across industry, (b) fast-tracking scientific innovation within AI R&D of explainability, and (c) implementing robust end-user education. I end with recommendations for how to advance such a collaborative effort.

2 Case Study: Autonomous Vehicle (AV) Technology

Until April of this year, the auto industry (and all other industries intending to or already making use of AI-enabled technology) had arrived at a bifurcated road: competitively move forward business as usual, or collectively step on the brakes, critically evaluate the capabilities of current AV technology, and set ethical and sustainable long-term goals. Now, designing, developing, and deploying AI-enabled technology is no longer a free-for-all. The United States’ Federal Trade Commission (FTC) and the European Commission have both separately voiced the potential harms of AI and proposed ways to govern AI through legal means [5, 6]. Crucially, the FTC has warned to go after unfair and deceptive practice within the AI industry in the United States, and the European Commission has proposed legislation that, if passed, would create significant obligations and limitations on the use of AI by the member states of the European Union. Specific to AVs, regulators in the United States now require manufacturers and operators to report incidents involving their driver-assistance and automated driving systems within one day of learning of a crash [15]. Leaving aside the FTC’s and the European Commission’s very different approaches to the governance of AI, the bottom line is clear: theoretical best practices are inadequate as the stakes are now higher than ever under the pressure of legal requirements to assure autonomy from these systems.

As with any technological change and the challenges that arise prior to its full-fledged adoption, these new demands can either hamper innovation or spur it. The first claim of this paper is that the science of assured autonomy is gathering unparalleled momentum outside of the engineering world for a new era of growth, and it crucially necessitates a human-centered approach. Assured autonomy is contingent on two

factors: (1) the system can accomplish goals independently, or with minimal supervision from human operators in environments that are complex and unpredictable, and (2) the system's capability is guaranteed and thus safe, secure, predictable, and reliable. These two factors underscore, moreover, two essential requirements: (i) AI-based systems with human-like intelligence capacities to indeed navigate the world as efficiently as we humans do, and (ii) explainability, or transparency, in how automated decisions are made to provide a level of interpretability for actions similar to how we humans provide reasons behind our own actions. While much work is being done to provide solutions to these two requirements, the science is incomplete [16] and needs alternative AI methods that can combine accuracy with transparency and, further yet, privacy concerns and mitigation [17]. Despite such recognizable considerations, the roadway to automation has not fully addressed them.

2.1 Reality vs. Fiction

Between misleading statements of an AV takeover [18], report after report of real-life problems with AV technology [19–22], and calls for honest discussions on the reality of AV capabilities and AI at large [23–27], the entire ecosystem around designing, developing, and deploying AVs has ignored the very core of the entire enterprise: humanity.

The concept of 'humanity' may well be intuitive, but I define it here in this context for the sake of clarity: that which fundamentally characterizes and defines us as a human in comparison to a machine. Instinctively, this refers to our intelligent ability to merge past experiences and common sense knowledge to think, feel, act, plan, make decisions, and problem-solve as we adapt to changing environments. These cognitive behavioral activities are at play in the case of driving a vehicle, or being around a vehicle. For example, when moving from point A to point B we make micro decisions in accordance with what is occurring in real time and what we know about roadways and drivers more broadly to navigate the world as successfully as possible. Acknowledging that what is being developed, i.e. machines, is not independent of the environment and has societal consequences, AVs must be able to accurately identify us, predict our behavior, and interact with us with relative ease. More bluntly, we humans are part of the engineering design equation because we coexist with roadways as drivers, construction workers, cyclists, jaywalkers, pedestrians, traffic guards, vendors, etc. We reflect our desires and goals through actions, and those actions are an integral part of what roadways entail and how they work, both through successes and failures. Any machine that enters today's roadways will face a plethora of contexts. Remove us and create a 100% interconnected robotized world –immune from hacking included– and we are not part of the engineering design equation.

Imagine the following scenario:

A maze of concrete, steel, and glass dominated. Gates opened and closed in synchrony: opening just one second before a fully automated driverless level 5 pod-like structure arrived, and remaining open for exactly the time it took to enter into the pod before locking into place as the pod smoothly pulled away. Gone were the speed limit signs; gone were the speed bumps; gone were the traffic lights; gone were the bicycle lanes; gone were the pedestrian walkways; gone

were the parking spaces; gone were all the penalties for violating traffic laws. Living beings were prohibited from entering “the vehicle zone,” as it was legally known.

Bridges were erected at every block to connect one side of the street to the other, every new bridge painstakingly merged to the previous one. In fact, so many bridges had been built and so many merged with swaths of steel and concrete that an entire floor exclusively for humans and animals had been created. The climate controlled and noiseless vehicle zone was a world of its own, an ever-growing cocoon impervious to unknowns. What the metro was to the humid underground, connected fully automated driverless level 5 vehicles were to the ground floor, and all living organic beings were to the scorching second floor.

You paid to enter an elevator or use the flights of stairs to move downwards for public transit or upwards for freedom...

Thus begins a short science fiction story I wrote as I read bombastic announcements of imminent deployment of AVs as early as 2019 [26, 27] and narrated to an audience of AV engineers and company executives during a keynote address in 2020 [14]. I re-share this fictionalized account of a future possibility to not only underscore the fantastical nature of a roadway devoid of humans, but to pose again the more fundamental question the AV industry needs to address: what do we as a society want to create with AVs? I ask in earnest as the industry must finally prioritize the unexpected complexities and challenges of real-life human behaviors and the ethical weight of human lives at stake over fast profits.

For conceptual clarity, the question can be broken down further: Do we want to create a machine that thinks and acts like us? Or do we want a machine that thinks and acts like us but in a better way in order to replace us? And what does ‘better’ even mean? Safer? How do we define safety? Do we want to replace our ambiguous, biased, distracted, emotional, error-prone, rule-breaking, and unpredictable behaviors? Is elimination of all human cognitive behavioral characteristics, or some of them, equivalent to engineering safety? Or do we want to create an entirely other thing? Would this other thing be not to replace us per se but to complement us in some way? And complement us in what way exactly? As a real-time guide that monitors our thoughts and behaviors moment-by-moment? Like a moral barometer that decides with or for us what is right or wrong? Or maybe we want the fictional world above where robo-vehicles coexist with each other in unwavering connected synchrony, free of interference from organic beings like ourselves? Or perhaps what we need most imminently is to create something that can interface with us much in the same way as we interface with our fellow human beings...

Whether addressing the external or internal environment of the vehicle, or both, the reality is that animals, the weather, and we humans are not going anywhere any time soon. Moreover, the physical infrastructure described in the story above is nowhere near actualization. We humans are, in effect, at the center of the problem. And all the above questions are not answerable by any one individual. Multiple stakeholders must address them collaboratively. Interdisciplinary and cross-disciplinary solutions are required because the problem is an interconnected human-machine-society issue with its resulting web of intertwined implications the science of engineering alone is not tackling. The characteristic ‘make x to yield y’ mindset of engineering systems to solve a ‘single’ problem is insufficient because context can neither be removed nor simply ignored. In the case of AVs, where the human sits in regards to the machine’s perspective –as a driver, passive occupant, or pattern of pixelated points in its path– is

critical to designing a predictable and thus reliable system to be used by, for, and around humans. The second claim of this paper is that AI has not been contextualized as belonging to the greater socio-technical ecosystem that science, technology, and society as a concept of dynamic interrelationships embody. The result has been a gap between paying heed to the values and choices of the people for which these AI-enabled systems are to be used by and the designing, developing, and deploying of AVs under the aura of vibrant change and innovation.

2.2 The Human-Centered Argument Distilled

Addressing the design of AVs from a human-centered perspective brings to the forefront two themes: (i) function allocation and (ii) human brain-inspired computing. Function allocation refers to the division of responsibility between humans and machines. In other words, who/what can and/or should do what and when and why. This is a decades old question, taking flight in the 1950s with the founding of the discipline of Human Factors as a way to directly address human problems within air navigation and traffic control. Specifically, it was "...a way of formulating a long-range integrated plan for human engineering research to parallel and support long range planning for equipment and systems design." [28, p. iii] Understanding the abilities, possibilities, and responsibilities of humans and/vs. machines, and translating that understanding into the design of machines ensures smooth interaction between users and the technology [25].

As performance demand rises for more intelligent and human-like artificial systems – most significantly with speech recognition and image classification capacity– human cognition-inspired models are becoming more and more invaluable. Human brain-inspired computing refers to the building of algorithms and architectures that mimic the natural forms of human cognition and the physiology of the human brain. This approach offers benefit for both the advancement of AI and the enlightenment of our understanding of human behavior. This method, like function allocation, is also decades old. Principally heralded by the creation of the first AI program in 1956 [29], the Logic Theorist was specifically built to resemble the problem-solving and decision-making skills of a human; it was capable of proving theorems in symbolic logic. Fast-forward to today and the need to surpass domain specificity and a reliance on vast numbers of high quality labeled training data is growing. The human mind/brain stands as a powerful example of maximal efficiency (i.e. domain generality/knowledge transference/inferential learning capability) within a finite space. Robots engineered in this way have illustrated improved performance [e.g. 30] and suggest promise for applications requiring power efficiency and cognitive abilities similar to that of humans.

Identifying and integrating the role of the human is inevitable for AV advancement. If the argument is to remove the human altogether and/or not replicate human cognition and behavior in the name of safety, e.g. to reduce the number of fatalities because of traffic accidents caused by human error, contrastive empirical support is needed on both sides regarding when and why humans succeed and fail, as well as when and why machines succeed and fail. Crash data between conventional vehicles and AVs, for example, are a stark reminder that minimization of error as an optimization objective for machine learning models is not a clear-cut metric for safety [25, 31]. On the other

hand, if the argument is to replicate human cognition and behavior in the name of safety, e.g. to increase the potential for seamless integration of human-machine interaction, a new empirical challenge needs to be prioritized over the traditional optimization objective of most machine learning models that are trained on data sets and deployed into the real world [See 24 for a unique cognitive behavioral parallelism between creative thinking and doing in the arts with driving a car in the city].

This assertion informs the third claim of this paper: insights from human perception and cognition as they relate to learning and adapting to ever-changing environments have an essential role in creating machine learning problem formations that are perfectly matched to the complex real-world tasks they will need to solve. In short, innovative AI R&D methods are urgently needed.

2.3 Explainability

Function allocation and human brain-inspired computing are not only critical to improving, for example, seamless interaction between humans and machines and accuracy of image classification systems [25], but they have further utility in the area of explainability. Explainability, explicability, interpretability, or transparency –all terms with varying definitions and periodic interchangeable usage [16, 32]– “deals with the capability to provide the human with understandable and relevant information on how an AI/[machine learning] ML application is coming to its results.” [16, p. 52] Again, human perception and cognition is our model for how we expect relationships in our world to function and how we test and explain the black box that is our mind/brain. The reasons we provide behind decision-making matter in our every day interactions; they serve as answers to agree or contest with on why a particular action or set of actions was made under a given situation. When presented with ‘why did you do x?’ we explain by breaking down what we believe to be the logic behind our decision(s). In the AI context where models are making predictions from orders of magnitude more data than any human being, the rationale behind the system’s output behavior also needs to be provided. Biases, abuses, and failures of these systems need explanation. Transparency of actions through explanation is important because it can strengthen the perception of honesty and improve trust.

In effect, we are at a pivotal moment in the history of AI-enabled technology and the development of legal constraints where the possible claim ‘the neural network we don’t understand is at fault’ by a manufacturer’s AV that crashed while on autopilot, for example, is insufficient. Moreover, explanations provided need to be understandable by a range of stakeholders that may include regulators, system engineers, system operators, and accident investigators with varying degrees of knowledge about data, machine learning, computations, algorithms, and the like. The following in Table 1 is a summary of the three types of explanations with increasing specificity for an AI-enabled system proposed by [16]. I include a sample of questions the three types could be utilized to answer.

While such types of explanations are a helpful starting point to setting standards for explainability, they have elicited varying degrees of transparency (e.g. different or

Table 1. Three proposed types of AI systems' explanations.

Type	Definition	Sample questions to answer
Simulatability	System/model	How does the system work?
Decomposability	System's components (e.g. model parameters, inputs, computations)	Where is the system's bias coming from? What is determining the system's safety criteria?
Algorithmic transparency	System's training algorithm	How is the system using the input data?

conflicting definitions between research groups; techniques are unique to specific architectures and thus non-generalizable) and incomplete adherence to the suggested types (e.g. some techniques explain the data but not the model and vice versa) across the board [16]. Moreover, even a battery of qualitative and quantitative tests on system-level methods targeted at the composition of the utilized neural network model do not necessarily provide insights on how the model functions and only inspire a false sense of confidence [16]. These observations underscore the current problems with explainability and the due diligence of such: (1) there is no academic and/or industry-wide unity on the various elements of explainability, and (2) the science behind machine learning must be supplemented by other techniques if any significant advancement with AI is to be made. Identification of principles and priorities such as 'transparency and explainability of AI systems' and 'accountability and responsibility' underpinning proposed legal frameworks [33] will be deficient without prioritizing and reframing the goals of the above two problems. The fourth claim of this paper is that consensus-based standards and compliance checks across industry and academic research labs are needed to significantly move forward with the potential of explainability as a means for assuring autonomy.

2.4 AV Automation Levels

In the meantime, revisiting SAE's levels of driving automation in Table 2 and narrowing in on agent responsibility underscores the indispensable requirement of keeping the human in the loop as the role of control gradually shifts from the human to the machine. Fundamental to this role shift is the diminishing (yet still expected) cognitive behavioral capacity required of the human as the machine takes over at level 3 and beyond [34]. Although level 3 is a turning point in the machine's monitoring capacity, there is still a noted reliance on critical human input. This is no easy feat when research reveals that human operator alertness and overall understanding of the traffic context and the system's functional limitations, among other things, are critical for successful decision-making and task takeover in emergencies [28], [35–36]. It is unfair to assume a human, dozing off in their automated driverless pod, for example, would have to be awakened and forced to intervene in a split-second emergency because the AV's sensors were unable to correctly classify and predict the behavioral trajectory of whatever was the cause of the resulting collision.

Not included in the table is the hypothesized number of vehicles from each level to be simultaneously on the road. The real world is not cleanly divided into strict cate-

Table 2. SAE levels of driving automation and their respective human-machine relationship.

Availability ^a	Level	Automation	Agent responsibility
On public roadways	0	None	Human driver (fully engaged)
On public roadways	1	Assisted	Human driver (fully engaged) with feet off; machine handles a function or two
On public roadways	2	Partial	Human driver (fully engaged) with feet and hands off; machine handles several functions
In closed course testing and on limited roadways	3	Conditional	Human (fully engaged enough to take control with notice) with feet, hands, and eyes off; machine handles most functions and monitors the environment under certain circumstances
In closed course testing and on limited roadways	4	High	Human (unengaged) with the option to take control; machine handles all functions and monitors the environment in certain circumstances
In closed course testing	5	Full	Machine handles all functions and monitors the environment; human is only a passenger and has no option to take control

^aAvailability of particular AV automation levels is a dynamic variable that varies across states here in the United States (and internationally) [37].

gories and boundaries. This is essential to consider and thus test because the very premise of the humanity argument presented here and in [25] is founded on a reality most likely before us: conventional vehicles (level 0), advanced driver assisted systems (ADAS) (levels 1 and 2), automated driving systems (ADS) (levels 3, 4, and 5), and everything else common to roadways will be eventually sharing roadways. Again, if the goal is to create a version of the fictional account presented earlier or simply a human-less roadway, then humanity takes on a distinct role than the one presented at length in [25] and discussed here. But if the goal is yet unclear as particular levels of AVs enter roadways beyond closed-course sunny areas with low speed limits, and a genuinely coordinated understanding and integration of the science among all stakeholders is pursued, let alone regaining trust from the general public, we have a moral obligation to keep humanity at the center of our AI building actions.

Fragile human-machine automation architectures need to be made robust. Without the general public's knowledge of what AVs realistically can and cannot do, the expectation is that 'if the car is on autopilot, it drives on its own.' The information feeding this expectation needs to be honest and deceptive tactics are not the answer. Possible claims of 'the user of the system was being inattentive' or 'the user didn't read the fine print in the car manual' by a manufacturer's AV that crashed while on

autopilot, for example, are insufficient. The fifth claim of this paper is that proper training and education for the end-user must be mandatory for any deployment of AI-enabled technology into the public domain.

2.5 The Trolley Problem is a Factor but not the Only Factor

Earlier I mentioned the indispensable role of trust in supporting cooperative relationships between humans. I then highlighted the importance of internal explainability as a tool for transparency of a system's outward actions. Linking trust and explainability is decision-making. We form opinions and choose actions via mental processes that are influenced by biases, reason, emotions, and memories. If assured autonomy is ultimately about guaranteeing a system's safety, security, predictability, and reliability, unpacking the logic behind the decision-making of all tasks performed by the system will be inevitable.

A decision-making situation that receives much attention is the classic Trolley Problem. The gruesome hypothetical is designed to test our moral intuitions in regards to choice making and the value we put on our decisions and the worth we give to others' lives. The problem generally states: a trolley is moving along in its tracks. Not too far ahead there are five workers lying in its direct path. On an alternate track there is only one worker. By chance, you happen to be next to a switch that can change the trolley's fate. If you pull the switch, the trolley will veer onto the alternative track and kill the one worker in its path. Question: do you pull the switch for the trolley to kill one person or do you leave as is and allow the trolley to kill five? There is no satisfactory right or wrong answer. The answer depends on a multitude of factors and conditions influenced by not only the environmental context per se of the moment and whether there is realistically any time to react with full awareness and judgment capacity, but by beliefs of self as a determiner of outcomes and the worth of others' lives. That is, who is to judge whose life is more valuable than another's? Such belief systems are not uniform across people and vary significantly across cultures [38].

This thought experiment, and subsequent transformations, calls urgent attention to the fact that we human drivers are routinely faced with a range of different moral decisions relating to our behavior in respect to other road users. Moreover, there is no one definitive answer to which decisions we want to delegate, and why, to AVs. I point out this decision-making situation among a host of possible decision-making situations AVs have to perform, let alone AI-enabled technologies more broadly, because it accentuates the very human intelligence reality we must contend with and it leads to the sixth claim of this paper: explaining automated decision-making may well remain a black box issue until we definitely answer the black box of how our own mind/brain learns, adapts, and executes new tasks effectively.

3 Our Ethical Responsibility Moving Forward

From where is our moral obligation borne to keep humanity at the center of our actions as it concerns AI? Two interrelated elements, one grander in its appeal and the other specific to the physical production of AI-enabled systems: human rights and business

optics. Every human being possesses basic rights and freedoms that need to be respected, protected, and fulfilled, and the longevity of businesses in democratic societies arguably depends on providing products and services that uphold our basic human rights and freedoms. The problem of assured autonomy is the 21st century's ethical and legal quandary as a result of the innovations from the fourth industrial revolution we are currently living.

2020 was coincidentally a year of clarity amongst the eruption of many entrenched problems within society, and the disintegration of trust was one of many consequences to boil to the surface. Publicly acknowledged perils of AI-enabled technologies fueled the already burning fire.

Whether considering intelligent adaptive capacities, human-to-machine takeover, and/or the ethical context of who, what, when, and why behind actions and outcomes, the Human is the core model from which to build a fruitful AI-enabled future. Table 3 summarizes the claims made throughout this paper as they were informed by a human-centered approach. These claims reveal the imperative role such an approach has on

Table 3. Summary of claims as informed by a human-centered approach.

Claim	Statement
1	The science of assured autonomy is entering a new era of growth
2	AI has not been contextualized as a socio-technical innovation
3	Alternative AI R&D methods that integrate insights from human perception and cognition must be prioritized
4	Industry and academic research labs must unify across standards and compliance measures of explainability
5	Training and education for the end-user are critical
6	Explaining the black box of automated decision-making rests on definitely answering the black box of our own mind/brain

building a sustainable foundation for the long-term development of assured autonomy because they endorse an interdependent and dynamic relationship whereby standards and compliance measures are collaboratively set, empirical paradigms within AI R&D of explainability are reframed, and the requirements of end-user education are shifted accordingly. Technology policy does not have to lag behind technology development.

3.1 Recommendations

Science does not exist in a vacuum; it is a social journey of inquiry where experiments are built and data acquired from the very multitude of experiences that shape our lives. As we humans are the major deterrent to the operational advancement of AVs and other AI-enabled technologies that require human-like intelligence to efficiently function within our human-centered world, productive solutions for assuring autonomy must be formulated via interdisciplinary, collaborative means. I propose the following recommendations by claim.

Claim 1: As mentioned, the science of assured autonomy will not significantly advance under an exclusive ‘build x to do y’ mindset. AI-enabled technologies are complex and evolving, dynamic systems entering into a complex and evolving, dynamic world. The successful building of ‘x’ to do ‘y’ requires recognizing that a range of interconnected factors and an array of possible consequences surround ‘x’ and ‘y’. To not recognize and act upon the interconnected factors and consequences AI-enabled (and non-AI-enabled) systems have beyond their individual physical and computational parts is to remain bound to intellectual confines no longer tenable by today’s changing legal landscape, let alone the justifiable uproar of voices demanding societal reforms. Safety, security, predictability, and reliability are real-world requirements not necessarily critical within static, closed-course laboratory environments. Furthermore, they are simply not answerable through purely mathematical means. This argues forcefully that a paradigm shift is needed within STEM education and STEM workforce industry culture to prepare and support our AI builders of today and tomorrow. Specifically, an interdisciplinary approach to learning and making must be instituted that connects multiple disciplines outside of traditional STEM fields and incentivizes their integration [e.g. 39]. Higher education learning institutions with AI programs can start by boldly reorienting their curricular options and evaluation metrics. Industry and government research sponsors can also start by requiring researchers to consider the societal ramifications of their work and provide real-time mitigation strategies.

Claim 2: The above can be extended to AI more broadly. In effect, this claim calls for an even broader overhaul across education and workforce culture that not only impacts the STEM fields and their respective industries, but the social sciences, humanities, and arts as well. Again, AI is not just an engineering problem; it’s a societal matter that intersects with communities and all the unique individuals that define those communities. Everyone needs to thoroughly understand what AI is so that engagement with the technology can be meaningful, purposeful, and equitable across the entire AI lifecycle of design, development, and deployment. Foundational knowledge of data, machine learning, computations, and algorithms is just as important as foundational knowledge of legal frameworks, ethics, and storytelling techniques. Companies need to spend more aggressively in workforce development both in training their own employees and evaluating their progress, and in attracting new talent through pre-college and college internships and mid-career training fellowships and other work opportunities. Research funding agencies could also create targeted programs for the development of cross-disciplinary AI talent.

Claim 3: Typical train-then-deploy machine learning systems are increasingly failing to improve the intelligence capacities of AI-enabled systems. Risky, outside-of-the-box research proposals can no longer fall under the curiosity-driven funding category; they need to be the mainstream. We need computational models that accurately mimic our capacity to constantly compare, contrast, and collate new information from actions performed by us and by those of others and innovative, cross-disciplinary research projects have the best chance at discovering a holistic solution. The United States needs a science and technology strategy for AI R&D that is unabashedly risk favorable.

Claim 4: Industry needs greater cooperation among its peers. At the macro level, this is a national security issue with transnational implications. Who decides the narrative of what AI is, can, and should be, pursues a specific science and technology strategy, and accomplishes goals will be the leading influencer. Given the reality of how globally networked scientific and engineering capabilities and innovation processes are today, a collaborative approach is requisite. This is where a robust network of international organizations with clearly defined standards and regulations that uphold democratic values and universal rights become paramount in pressuring companies to adhere for competitive advantage: the OECD Principles on AI [40], RAI Certification [41], the Global Partnership on AI (GPAI) [42], and others.

Claim 5: Following from the above and moving to the micro level, the end-user needs to be well informed, not oblivious to the realities of today's AI capacities. While all the above recommendations apply here as well, industry peers can unite now around end-user education policies in the form, for example, of uniform educational campaigns, and the compliance of such. This type of transparency even simplifies oversight and offers clear entrance points for revision, if and when needed.

Claim 6: We humans are the elephant in the room and our mind/brain holds the key to what we have the potential to build, both from a metacognitive sense and a technical sense. Priority must lie in reverse engineering the human mind/brain and pausing to reflect as a human species on what we want with AI in this next phase in our evolutionary history.

3.2 Beyond AVs

While I have used AVs to contextualize the claims made and summarized in Table 3, they are by no means the end-all of AI-enabled technologies. In effect, AVs are but one example of many established and burgeoning AI-enabled systems resulting from the advancement of sensors, software, and emerging technologies that constitute the Internet of Things (IoT). From empathetic AI and companion robots to drones and urban air mobility, the list is long. Assured autonomy becomes an even more critical issue by the minute as AI-enabled use cases amplify, increasing the network of connected smart systems and thus the number of interdependent factors involved and possible outcomes of failure. But whether we build one AI-enabled technology or $N \geq 1$ AI-enabled technologies the problem remains: it or they will be used by, for, and around humans.

The question now is: can we succeed in uniting efficiently and understanding ourselves better enough to create smarter machines not yet fathomable by our own intelligence?

Acknowledgements. Thanks to Dr. I. Gonzalez for her insightful comments and to the Institute for Human Intelligence for providing the required resources for this research.

References

1. Leslie, D.: Understanding bias in facial recognition technologies: an explainer. The Alan Turing Institute (2020). <https://doi.org/10.5281/zenodo>
2. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J.: How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**(4), 582–584 (2021)
3. Lee, P.: Learning from Tay’s introduction. Microsoft – Official Microsoft Blog, 25 March 2016. <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>
4. Tesla Deaths: <https://www.tesladeaths.com/>. Accessed 24 June 2021
5. Jillson, E.: Aiming for truth, fairness, and equity in your company’s use of AI. Federal Trade Commission – Business Blog, 19 April 2021. <https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>
6. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. European Commission, April 2021. <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>
7. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization, June 2021. Licence: CC BY-NC-SA 3.0 IGO. https://www.who.int/publications/i/item/9789240029200?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axioswhatsnext&stream=science
8. The Association for Unmanned Vehicle Systems International (AUVSI) Industry News Webpage. <https://www.auvsi.org/news>. Accessed 24 June 2021
9. Trust in tech is wavering and companies must act. Edelman Research, April 2019. <https://www.edelman.com/research/2019-trust-tech-wavering-companies-must-act>
10. Rainie, L., Anderson, J., Vogels, E.A.: Experts doubt ethical AI design will be broadly adopted as the norm within the next decade. Pew Research Center, June 2021. <https://www.pewresearch.org/internet/2021/06/16/experts-doubt-ethical-ai-design-will-be-broadly-adopted-as-the-norm-within-the-next-decade/>
11. 2021 Edelman Trust Barometer. Edelman Research, January 2021. <https://www.edelman.com/trust/2021-trust-barometer>
12. Erikson, E.: *Childhood and Society*. Norton & Company Inc, New York (1950)
13. Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C.: Not so different after all: a cross-discipline view of trust. *Acad. Manag. Rev.* **23**(3), 393–404 (1998)
14. López-González, M.: Regaining sight of humanity on the roadway to automation. In: IS&T International Symposium on Electronic Imaging: Autonomous Vehicles and Machines, IS&T. Springfield, Virginia (2020). <https://doi.org/10.2352/ISSN.2470-1173.2020.16.AVM-088>.
15. National Highway Traffic Safety Administration (NHTSA): NHTSA orders crash reporting for vehicles equipped with advanced driver assistance systems and automated driving systems. NHTSA Press Release, 29 June 2021. <https://www.nhtsa.gov/press-releases/nhtsa-orders-crash-reporting-vehicles-equipped-advanced-driver-assistance-systems>
16. EASA and Daedalean: Concepts of design assurance for neural networks (CoDANN) II, May 2021. <https://daedalean.ai/tpost/kg4j07xlx1-daedalean-and-easa-conclude-second-proje>
17. General Data Protection Regulation (GDPR). <https://gdpr.eu/tag/gdpr/>. Accessed 13 July 2021
18. Waldrop, M.M.: Autonomous vehicles: no drivers required. *Nat. News* **518**(7537), 21–22 (2015)

19. Miller, R.: Closing the curtains on safety theater, Medium, 18 April 2019. <https://medium.com/pronto-ai/closing-the-curtains-on-safety-theater-f442b70645a4>
20. Boudette, N.E.: Despite high hopes, self-driving cars are ‘way in the future’. The New York Times, 17 July 2019. <https://www.nytimes.com/2019/07/17/business/self-driving-autonomous-cars.html>
21. Taub, E.A.: How jaywalking could jam up the era of self-driving cars. The New York Times, 1 August 2019. <https://www.nytimes.com/2019/08/01/business/self-driving-cars-jaywalking.html>
22. Edmonds, E.: AAA warns pedestrian detection systems don’t work when needed most, AAA NewsRoom, 3 October 2019. <https://newsroom.aaa.com/2019/10/aaa-warns-pedestrian-detection-systems-dont-work-when-needed-most/>
23. Young, S.: The moral algorithm: how to set the moral compass for autonomous vehicles. Report produced by Gowling WLG (UK), LLC, December 2016
24. López-González, M.: Theoretically automated conversations: collaborative artistic creativity for autonomous machines. In: IS&T International Symposium on Electronic Imaging: Human Vision and Electronic Imaging, IS&T. Springfield, Virginia (2018). <https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-531>
25. López-González, M.: Today is to see and know: an argument and proposal for integrating human cognitive intelligence into autonomous vehicle perception. In: IS&T International Symposium on Electronic Imaging: Autonomous Vehicles and Machines, IS&T. Springfield, Virginia (2019). <https://doi.org/10.2352/ISSN.2470-1173.2019.15.AVM-054>
26. Naughton, N.: GM moves to deploy driverless car fleet in 2019, The Detroit News, 12 January 2018. <https://www.detroitnews.com/story/business/autos/general-motors/2018/01/12/gm-driverless-car-fleet-cruise-av/109381232/>
27. Walker, J.: The self-driving car timeline – predictions from the top 11 global automakers, Emerj, 21 December 2018. <https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/>
28. Fitts, P.M. (ed.): Human engineering for an effective air-navigation and traffic-control system. In: Report prepared for the Air Navigation Development Board. National Research Council, Washington, D.C., March 1951
29. Stefferud, E.: The logic theory machine: a model heuristic program. Memorandum RM-3731-CC, The Rand Corporation, Santa Monica, CA, June 1963
30. Lázaro-Gredilla, M., Lin, D., Guntupalli, J.S., George, D.: Beyond imitation: zero-shot task transfer on robots by learning concepts as cognitive programs. *Sci. Rob.* **4**(26), eaav3150 (2019)
31. Favarò, F.M., Nader, N., Eurich, S.O., Tripp, M., Varadaraju, N.: Examining accident reports involving autonomous vehicles in California. *PLoS ONE* **12**(9), e0184952 (2017)
32. Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* **16**(3), 31–57 (2018)
33. Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M., Briggs, M.: Artificial intelligence, human rights, democracy, and the rule of law: a primer. The Council of Europe (2021)
34. National Highway Traffic Safety Administration, United States Department of Transportation, Automated Vehicles for Safety. <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety>
35. Safety Alert for Operators (SAFO - 13002): Manual flight operations. U.S. Department of Transportation Federal Aviation Administration, 4 January 2013. https://www.faa.gov/other_visit/aviation_industry/airline_operators/airline_safety/safo/all_safo/media/2013/SAFO13002.pdf

36. National Transportation Safety Board: Collision between vehicle controlled by developmental automated driving system and pedestrian. Public Meeting of 19 November 2019. <https://www.nts.gov/news/events/Documents/2019-HWY18MH010-BMG-abstract.pdf>
37. AutoInsurance.org: Which states allow self-driving cars? (2021 Update), 26 February 2021. <https://www.autoinsurance.org/which-states-allow-automated-vehicles-to-drive-on-the-road/>. Accessed 30 June 2021
38. Awad, E., et al.: The moral machine experiment. *Nature* **563**(7729), 59–64 (2018)
39. López-González, M.: For female leaders of tomorrow: cultivate an interdisciplinary mindset. In *Women in Engineering (WIE) Forum USA East*. IEEE (2017). <https://doi.org/10.1109/WIE.2017.8285606>
40. Oecd.org: OECD Principles on Artificial Intelligence. <https://www.oecd.org/going-digital/ai/principles/>. Accessed 13 July 2021
41. Responsible AI Institute: RAI Certification. <https://www.responsible.ai/certification>. Accessed 13 July 2021
42. Gpai.org: The Global Partnership on Artificial Intelligence. <https://www.gpai.ai/>. Accessed 13 July 2021