



BY STUART E. MIDDLETON, EMMANUEL LETOUZÉ,
ALI HOSSAINI, AND ADRIANE CHAPMAN

Trust, Regulation, and Human- in-the-Loop AI within the European Region

ARTIFICIAL INTELLIGENCE (AI) systems employ learning algorithms that adapt to their users and environment, with learning either pre-trained or allowed to adapt during deployment. Because AI can optimize its behavior, a unit's factory model behavior can diverge after release, often at the perceived expense of safety, reliability, and human controllability. Since the Industrial Revolution, trust has ultimately resided in regulatory systems set up by governments and standards bodies. Research into human interactions with autonomous machines demonstrates a shift in the locus of trust: we must *trust* non-deterministic systems such as AI to self-regulate, albeit within boundaries. This radical shift is one of the biggest issues facing the deployment of AI in the European region.

Trust has no accepted definition, but Rousseau²⁸ defined it as “a psychological state comprising the

intention to accept vulnerability based upon positive expectations of the intentions or behavior of another.” Trust is an attitude that an agent will behave as expected and can be relied upon to reach its goal. Trust breaks down after an error or a misunderstanding between the agent and the trusting individual. The psychological state of trust in AI is an emergent property of a complex system, usually involving many cycles of design, training, deployment, measurement of performance, regulation, redesign, and retraining.

Trust matters, especially in critical sectors such as healthcare, defense, and security, where duty of care is foremost. Trustworthiness must be planned, rather than an afterthought. We can *trust in AI*, such as when a doctor uses algorithms to screen medical images.²⁰ We can also *trust with AI*, such as when journalists reference a social network algorithm to analyze sources of a news story.³⁷ Growing adoption of AI into institutional systems relies on citizens to trust in these systems and have confidence in the way these systems are designed and regulated.

Regional approaches for managing trust in AI have recently emerged, leading to different regulatory regimes in the U.S., the European region, and China. We review these regulatory divergences. Within the European region, research programs are examining how trust impacts user acceptance of AI. Examples include the UKRI Trustworthy Autonomous Systems Hub,^a the French Confiance.ai project,^b and the German AI Breakthrough Hub.^c Europe appears to be developing a “third way,” alongside the U.S. and China.¹⁹

Healthcare contains many examples of AI applications, including online harm risk identification,²⁴ mental health behavior classification,²⁹ and

a <https://www.tas.ac.uk>

b <https://www.confiance.ai>

c <https://breakthrough-hub.ai>



automated blood testing.²² In defense and security, examples include combat management systems⁹ and using machine learning to identify chemical and biological contamination.¹ There is a growing awareness within critical sectors^{15,33} that AI systems need to address a “public trust deficit” by adding reliability to the perception of AI. In the next two sections, we discuss research highlights around the key trends of building safer and more reliable AI systems to engender trust and put humans in the loop with regard to AI systems and teams. We conclude with a discussion about applications, and what we consider the future outlook is for this area.

Recent Changes in the Regulatory Landscape for AI

The E.U. is an early mover in the race to regulate AI, and with the draft E.U. AI Act,^d it has adopted an *assurance-based regulatory environment* using yet-to-be-defined AI assurance standards.

d <https://bit.ly/3FATnNj>

These regulations build upon GDPR data governance and map AI systems into four risk categories. The lowest risk categories self-regulate with transparency obligations. The highest risk categories require first-party or third-party assessments enforced by national authorities. Some applications are banned outright to protect individual rights and vulnerable groups.

The U.K. AI Council AI Roadmap^e outlines a sector-specific *audit-led regulatory environment*, along with principles for governance of AI systems including open data, AI audits, and FAIR (Findable, Accessible, Interoperable, Reusable) principles. An example of sector-specific governance is the U.K. online safety bill,^f which assigns a duty of care to online service providers and mandates formal risk assessments by the U.K. telecom regulator OFCOM.

Outside the European region, the U.S. National Security Commission on

e <https://www.gov.uk/government/publications/ai-roadmap>

f <https://www.gov.uk/government/publications/draft-online-safety-bill>

AI report 2021^g outlined a *market-led regulatory environment*, with government focus areas of robust and reliable AI, human-AI teaming, and a standards-led approach^h to testing, evaluation, and validation. China’s AI development plan²⁷ emphasizes societal responsibility; companies chosen by the Chinese state to be AI champions follow national strategic aims, and state institutions determine the ethical, privacy, and trust frameworks around AI.

The European region, driven by U.K. and E.U. AI regulation, is creating a “third way” alongside the AI regulation adopted by the U.S. and China. This “third way” is characterized by a strong European ethical stance around AI applications, for example limiting the autonomy of military AI systems, in direct contrast to China, where autonomy for AI-directed weapons is actively encouraged as part of its military-civil fusion strategy.¹⁴ It also is characterized by a strong European

g <https://www.nsc.gov/2021-final-report>

h <https://www.nist.gov>

The E.U. is an early mover in the race to regulate AI, and with the draft E.U. AI Act, it has adopted an assurance-based regulatory environment using yet-to-be-defined AI assurance standards.

focus on a citizen's right to data privacy and the limits set on secondary data processing by AI applications, in contrast to China and the U.S., where state-sponsored strategic aims or weak commercial self-regulation around AI applications frequently override data privacy concerns. An example of this "third way" in action is the European city of Vienna becoming the first city in the world to earn the IEEE AI Ethics Certification Mark,³⁰ which sets standards for transparency, accountability, algorithmic bias, and privacy of AI products. How different regional approaches to AI regulation perform in the heat of geo-political AI competition is likely to shape how regional AI research is conducted for many years to come.

Building Safe and Reliable AI to Engender Trust

Assuring safe, reliable AI systems can provide a pathway to trust. However, non-deterministic AI systems require more than just the application of quality assurance protocols designed for conventional software systems in well-regulated regions such as Europe. New methods are emerging for the assurance of the machine learning life cycle from data management to model learning and deployment.²

Exploratory data analysis and adversarial generative networks help assure training data comes from a trusted source, is fit for the purpose, and is unbiased. *Built-in test (BIT)* techniques support model deployment, such as watchdog timers or behavioral monitors, as well as "last safe" *model checkpointing* and *explainable AI* methods. Active research focuses on explainable machine learning.⁵ Approaches include *explanation by simplification*, such as local interpretable model-agnostic explanations (LIME) and counterfactual explanations; *feature relevance techniques*, such as Shapley Additive Explanations (SHAP) and analysis of random feature permutations; *contextual and visual explanation* methods such as sensitivity analysis and partial dependence plots; and full life-cycle approaches such as the use of provenance records. Research challenges for assurance of machine learning include detection of problems before critical failures, con-

tinuous assurance of adaptive models, and assessing levels of independence when multiple models are trained on common data.

The manufacturing sector and smart cities deployments increasingly are using digital twins,³⁶ simulations of operating environments, to provide pre-deployment assurance. Digital twins also are used in healthcare,⁸ for example to assure pre-surgical practice, and other critical sectors. A recent U.K.-hosted RUSI-TAS Conference³⁵ discussed how digital twins can provide AI models with a safe space to fail. Other research trends include probing vulnerabilities of AI to accidents or malicious use. This includes examining how malicious actors can exploit AI.¹¹ Attack vectors include adversarial inputs, data poisoning, and model stealing. Possible solutions include safety checklists¹² and analysis of hostile agents that use AI to subvert democracies.³¹

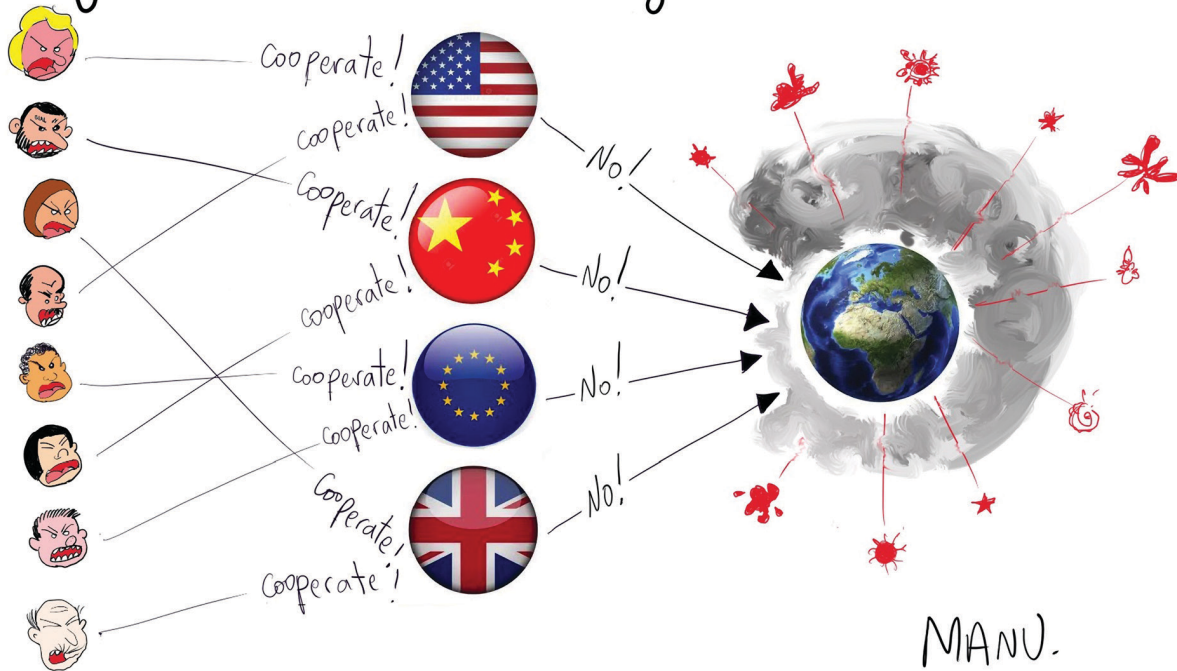
Safe and Reliable AI has received a lot of attention in the European region recently compared to the U.S. and China, and it is no coincidence that every one of the works cited in this section are from authors based in this region. This level of activity is probably motivated by the assurance and audit-based European regulatory stances. The more we understand the vulnerabilities and assurance protocols of AI, the safer and more reliable AI systems will become. Safe, transparent systems that address user concerns will encourage public trust.

Human and Society in the Loop

Human-in-the-Loop (HITL) systems are grounded in the belief that human-machine teams offer superior results, building trust by inserting human oversight into the AI life cycle. One example is when humans mark false positives in email spam filters. HITL enhances trust in AI by optimizing performance, augmenting data, and increasing safety. It enhances trust by providing transparency and accountability: unlike many deep learning systems, humans can explain their decisions in natural language.

However, the AI powering social media, commerce, and other activities may erode trust and even sow discord.⁴ If perceived as top-down oversight from

If the world were a giant neural network



experts, HITL is unlikely to address public trust deficits. Society-in-the-Loop (SITL) seeks broader consensus by extending HITL methods to larger demographics,^{16,25} for instance by crowdsourcing the ethics of autonomous vehicles to hundreds of thousands of people. Another approach is co-design with marginalized stakeholders. The same imperative drives CODES (Council for the Orientation of Development and Ethics) in AI and data-driven projects in developing countries,ⁱ where representatives of local stakeholder groups provide feedback during project life cycles. SITL combined with mass data literacy⁷ may reweave the fabric of human trust in and with AI.

A growing trend is to add humans into deep learning development and training cycles. Human stakeholders co-design AI algorithms to encourage responsible research innovation (RRI), embed end-user values, and consider the potential for misuse. During AI training, traditional methods such as *adversarial training* and *active learning* are applied to the deep learning models^{13,21} using humans to label

uncertain or subjective data points during training cycles. *Interactive sense making*¹⁷ and *explainable AI*⁵ also can enhance trust by visualizing AI outputs to reveal training bias, model error, and uncertainty.

Research into HITL is much more evenly spread across the European, U.S., and Chinese regions than work on safe and reliable AI, with about half the work cited in this section from authors based in the European region. Where the European region does differentiate itself is with a stronger focus on HITL to promote ethical AI and responsible innovation, as opposed to the U.S. and China, where there is a tighter focus on using HITL to increase AI performance.

Applications in Critical Sectors

AI offers considerable promise in the following sectors. Each illustrates high-risk, high-reward scenarios where trust is critical to public acceptance.

Defense. General Sir Patrick Sanders, head of U.K. Strategic Command, recently emphasized, “Even the best human operator cannot defend against multiple machines making thousands of maneuvers per second

at hypersonic speeds and orchestrated by AI across domains.”¹⁸ While human-machine teaming dominates much current military thinking, by taking humans *out* of the loop AI transforms the tempo of warfare beyond human capacity. From strategic missile strikes to tactical support for soldiers, AI impacts every military domain and, if an opponent has a high tolerance for error, it offers unstoppable advantages. Unless regulated by treaty, future warriors and their leaders will likely trust AI as a matter of necessity.

Law enforcement and security. Law enforcement is more nuanced. Though used only for warnings, Singapore’s police robots have provoked revulsion in European press,³⁴ and the E.U. AI Act reflects this attitude by classifying law enforcement as high-risk. Some groups have claimed ambiguities in the E.U. AI Act leave the door open for bias, unrestrained surveillance, and other abuses,³² but at minimum it provides a framework for informed progress while asserting the European region’s core values.

Healthcare. Healthcare interventions directly impact lives. Research

i <https://datapopalliance.org>

into diagnostic accuracy shows that AI can improve healthcare outcomes.^{6,10,23,26} However, starting with patients and physicians, trust cascades upward, and as Covid has shown, trust is ultimately political, and thus needs to be nurtured carefully.

Transportation. Self-driving cars may receive the most publicity, but AI also is applied to mass transit, shipping, and trucking. Transportation involves life-or-death decisions, and the introduction of AI is changing the character of liability and assurance. These questions reflect a fundamental question which is being debated today: Who does the public trust to safely operate a vehicle?

Future Outlook


We think future standards for assurance will need to address the non-deterministic nature of autonomous systems. Whether robotic or distributed, AI is effectively an entity, and regulation, management, and marketing will need to account for its capacity to change.

Many projects currently are exploring aspects of bringing humans into the loop for co-design and training of AI systems and human-machine teaming. We think this trend will continue, and if coupled with genuine transparency, especially around admitting AI mistakes and offering understandable explanations for why these mistakes happened, offers a credible pathway to improving the state of public trust in AI systems being deployed into society.

We think that increasingly, *Trust with AI* will shape how citizens trust information, which has the potential to reduce the negative impact of attempts to propagate disinformation. If citizen trust in the fabric of AI used within society is reduced, then *trust in AI* itself will weaken. This is likely to be a major challenge for our generation.

Creating regulatory environments that allow nation-states to gain commercial, military, and social advantages in the global AI race may be the defining geopolitical challenge of this century. Regulation around AI has been developing worldwide, moving from self-assessment guidelines³ to frameworks for national or transna-

tional regulation. We have noted that there are clear differences between the European region and other areas with robust capacity in AI, notably the need for public acceptance. The future will be a highly competitive environment, and regulation must balance the benefits of rapid deployment, the willingness of individuals to trust AI, and the value systems which underlie trust.

Acknowledgments. This work was supported by the Engineering and Physical Sciences Research Council (EP/V00784X/1), Natural Environment Research Council (NE/S015604/1), and Economic and Social Research Council (ES/V011278/1; ES/R003254/1). 

References

- Alan Turing Institute. Data Study Group Final Report: DSTL—Anthrax and nerve agent detector. (2021); <https://doi.org/10.5281/zenodo.4534218>
- Ashmore, R., Calinescu, R., and Paterson, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Comput. Surv.* 54, 5 (2021), Article 111; <https://doi.org/10.1145/3453444>
- Ayling, J. and Chapman, A. Putting AI ethics to work: Are the tools fit for purpose? *AI Ethics* (2021); <https://doi.org/10.1145/3453444>
- Barrett, P., Hendrix, J., and Sims, G. How tech platforms fuel U.S. political polarization and what government can do about it. The Brookings Institution 27 (Nov. 2021); <https://brook.gs/3sK3Cev>
- Belle, V., and Papantonis, I. Principles and Practice of Explainable Machine Learning. (2020); [arXiv:2009.11698](https://arxiv.org/abs/2009.11698)
- Bhandari, M., Zeffiro, T., and Reddiboina, M. Artificial intelligence and robotic surgery: Current perspective and future directions. *Curr Opin Urol.* 30, 1 (2020), 48–54; [doi:10.1097/MOU.0000000000000692](https://doi.org/10.1097/MOU.0000000000000692)
- Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., and Shoup, N. Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data, Data-Pop Alliance White Paper. Sept. 29, 2015; <https://bit.ly/3qNgBtm>
- Bruynseels, K., Santoni de Sio, F., and van den Hoven, J. Digital twins in health care: Ethical implications of an emerging engineering paradigm. *Front. Genet.* 9, 31 (2018); <https://doi.org/10.3389/fgene.2018.00031>
- AI and data science: Defense science and technology capability. Aug. 1, 2021; <https://www.gov.uk/guidance/ai-and-data-science-defence-science-and-technology-capability>
- Gumbs, A.A., Frigerio, I., Spolverato, G., Croner, R., Illanes, A., Chouillard, E., and Elyan, E. Artificial intelligence surgery: How do we get to autonomous actions in surgery? *Sensors (Basel)* 21, 16 (2021); <https://www.mdpi.com/1424-8220/21/16/5526>
- Hartmann, K., Steup, C. Hacking the AI—The next generation of hijacked systems. In *Proceedings of the 12th Intern. Conf. Cyber Conflict*, 2020, 327–349; [doi:10.23919/CyCon49761.2020.9131724](https://doi.org/10.23919/CyCon49761.2020.9131724)
- Hunt, E.R., and Hauert, S. A checklist for safe robot swarms. *Nature Machine Intelligence* (2020); [doi:10.1038/s42256-020-0213-2](https://doi.org/10.1038/s42256-020-0213-2)
- Kanchinadam T., Westpfahl, K., You, Q., and Fung, G. Rationale-based human-in-the-loop via supervised attention. In *Proceedings of 1st Workshop on Data Science with Human in the Loop* (Aug. 24, 2020); <https://bit.ly/3eYIKJA>
- Kania, E.B. Chinese military innovation in the AI revolution. *The RUSI J* 164, 5–6 (2019), 26–34; [doi:10.1080/03071847.2019.1693803](https://doi.org/10.1080/03071847.2019.1693803)
- Kerasidou, C., Kerasidou, A., Buscher, M., and Wilkinson, S. Before and beyond trust: Reliance in medical AI. *J Medical Ethics* (2021); <https://dx.doi.org/10.1136/medethics-2020-107095>
- Larsson, S. The socio-legal relevance of artificial intelligence. *Droit et société* 103, (2019) 573–593; [doi:10.3917/drs1.103.0573](https://doi.org/10.3917/drs1.103.0573)
- Middleton, S.E., Lavorgna, L., Neumann, G., and Whitehead, D. Information extraction from the long tail: A socio-technical AI approach for criminology investigations into the online illegal plant trade. *WebSci '20 Companion*, 2020.
- Ministry of Defense. Commander of Strategic Command RUSI conference speech, May 26, 2021; <https://www.gov.uk/government/speeches/commander-of-strategic-command-rusi-conference-speech>
- Morton, S. and Booth, M. The EU's "third way" to AI regulation. Pillsbury, Sept. 22, 2021; <https://www.internetandtechnologylaw.com/eu-third-way-ai-regulation/>
- NHS-X. Cancer digital playbook. (2021); <https://www.nhs.uk/key-tools-and-info/digital-playbooks/cancer-digital-playbook/>
- Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial NLI: A new benchmark for natural language understanding. *ACL* (2020)
- Pinpoint. Early Cancer Detection. (2021); <https://www.pinpointdatascience.com>
- Prabhakar, B., Singh, R.K., and Yadav, K.S. Artificial intelligence (AI) impacting diagnosis of glaucoma and understanding the regulatory aspects of AI-based software as medical device. *Computerized Medical Imaging and Graphics* 87 (2021).
- ProTechThem. ESRC grant ES/V011278/1 (2021); <http://www.protechthem.org>
- Rahwan, I. Society-in-the-loop: programming the algorithmic social contract. *Ethics Inf Technol* 20, (2018), 5–14; [doi:10.1007/s10676-017-9430-8](https://doi.org/10.1007/s10676-017-9430-8)
- Rangarajan, A.K., Ramachandran, H.K. A preliminary analysis of AI-based smartphone application for diagnosis of COVID-19 using chest X-ray image. *Expert Systems with Applications* 183, (2021); [doi:10.1016/j.eswa.2021.115401](https://doi.org/10.1016/j.eswa.2021.115401)
- Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., and Floridi, L. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & Soc* 36 (2021) 59–77; [doi:10.1007/s00146-020-00992-2](https://doi.org/10.1007/s00146-020-00992-2)
- Rousseau, D.M., Sitkin, S.B., Burt, R.S., and Camerer, C. Not so different after all: A cross-discipline view of trust. *Academy of Management Rev.* 23, (1998), 393–404; [doi:10.5465/AMR.1998.926617](https://doi.org/10.5465/AMR.1998.926617)
- SafeSpacesNLP. UKRI TAS agile project, (2021); <https://www.tas.ac.uk/safespacesnlp>
- Schabus, D. The IEEE CertifAIEd Framework for AI Ethics Applied to the City of Vienna. IEEE Standards Assoc. (2021); <https://bit.ly/3EZWzRp>
- Schia, N.N. and Gjesvik, L. Hacking democracy: managing influence campaigns and disinformation in the digital age. *J. Cyber Policy* 5, 3 (2020), 413–428; [doi:10.1080/23738871.2020.1820060](https://doi.org/10.1080/23738871.2020.1820060)
- Skelton, S.K. NGO Fair Trials calls on E.U. to ban predictive policing systems. *ComputerWeekly* (2021); <https://bit.ly/3mG4uWV>
- Taddeo, M., McCutcheon, T., and Floridi, L. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nat Mach Intell* 1, (2019), 557–560; [doi:10.1038/s42256-019-0109-1](https://doi.org/10.1038/s42256-019-0109-1)
- The Guardian via Agence France Presse. 'Dystopian world': Singapore patrol robots stoke fears of surveillance state; <https://bit.ly/3EBQRou>
- RUSI-TAS. Trusting Machines? *RUSI-TAS 2021 Conf.*; <https://www.tas.ac.uk/eventslist/trusting-machines/trust-machines-conference-programme/>
- van der Valk, H., Haße, H., Möller, F., Arbter, M., Henning, J., and Otto, B. A Taxonomy of digital twins. In *Proceedings of AMCIS 2020*; <https://bit.ly/3HE6ZYI>
- European Commission, Horizon 2020 grant agreement 825297 (2021); <https://cordis.europa.eu/project/id/825297>

Stuart E. Middleton is a lecturer in computer science at the University of Southampton, Southampton, UK.

Emmanuel Letouzé is Marie Curie Fellow at Universitat Pompeu Fabra, Barcelona, Spain.

Ali Hossaini is Senior Visiting Research Fellow at Kings College London, UK.

Adriane Chapman is a professor in computer science at the University of Southampton, Southampton, UK.