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

By Stuart E. Middleton*, Emmanuel Letouzé, Ali Hossaini, Adriane Chapman

*Contact author: sem03@soton.ac.uk

Introduction

Artificial intelligence (AI) systems employ learning algorithms which adapt to their users and environment, with learning either pre-trained or allowed to adapt during deployment. Because AI can optimize its behaviour, a unit's factory model behaviour 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 [Rousseau 1998] define it as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour 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, defence & 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 [NHS-X 2021]. We can also *trust with AI*, such as when journalists reference a social network algorithm to analyse sources of a news story [WeVerify 2021]. 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 United States, 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[1], the French Confidiance.ai project[2] and the German AI Breakthrough Hub[3]. Europe appears to be developing a "third way" alongside the United States and China [Morton 2021].

Healthcare contains many examples of AI applications including online harm risk identification [ProTechThem 2021], mental health behaviour classification [SafeSpacesNLP 2021] and automated blood testing [Pinpoint 2021]. In defence and security, examples include combat management systems [DSTL 2021] and using machine learning to identify chemical and biological contamination [Alan Turing Institute 2021]. There is a growing awareness within critical sectors [Kerasidou 2020] [Taddeo 2019] that AI systems need to address a "public trust deficit" by adding reliability into 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 putting humans in the loop with

[1] https://www.tas.ac.uk

[2] https://www.confiance.ai

[3] https://breakthrough-hub.ai

regards AI systems and teams. We conclude with a discussion about applications and what we consider the future outlook is in this area.

Recent changes in regulatory landscape for AI

The EU is an early mover in the race to regulate AI, and with the draft EU AI Act[4] it has adopted an *assurance-based regulatory environment* using yet to be defined AI assurance standards. 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 UK AI Council AI Roadmap[5] outlines a sector-specific *audit-led regulatory environment*, along with principles for governance of AI systems including open data, AI audits and FAIR principles (Finable, Accessible, Interoperable, Reusable). An example of sector-specific governance is the UK online safety bill[6], which assigns a duty of care to online service providers and mandates formal risk assessments by the UK 's telecom regulator OFCOM.

Outside the European region the US National Security Commission on AI report 2021[7] outlined a *market-led regulatory environment*, with government focus areas of robust and reliable AI, human-AI teaming and a standards-led approach[8] to testing, evaluation, and validation. China's AI development plan [Roberts 2021] 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 UK and EU AI regulation, is creating a "third way" alongside the AI regulation adopted by the United States and China. This "third way" is characterised 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 [Kania 2019]. It is also characterised by a strong European 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 US 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 [Schabus 2021], 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

[4] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

[5] https://www.gov.uk/government/publications/ai-roadmap

[6] https://www.gov.uk/government/publications/draft-online-safety-bill

[7] https://www.nscai.gov/2021-final-report

[8] https://www.nist.gov

for the assurance of the machine learning lifecycle from data management to model learning and deployment [Ashmore 2021]. *Exploratory data analysis* and *adversarial generative networks* helps assure training data comes from a trusted source, is fit for purpose, and is unbiased. *Built-in test (BIT)* techniques support model deployment, such as watchdog timers or behavioural monitors, as well as "last safe" *model checkpointing* and *explainable AI* methods. Active research focusses on explainable machine learning [Belle 2020]. 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, partial dependence plots; and full lifecycle approaches such as the use of provenance records. Research challenges for assurance of machine learning include detection of problems before critical failures, continuous assurance of adaptive models, and assessing levels of independence when multiple models are trained on common data.

The manufacturing sector and smart cities deployments are increasingly using digital twins [van der Valk 2020], simulations of operating environments, to provide pre-deployment assurance. Digital twins are also used in healthcare [Bruynseels 2018], for example to assure pre-surgical practice, and other critical sectors. A recent UK-hosted RUSI-TAS Conference [RUSI-TAS 2021] 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 [Hartmann 2020]. Attack vectors include adversarial inputs, data poisoning, and model stealing. Possible solutions include safety checklists [Hunt 2020] and analysis of hostile agents that use AI to subvert democracies [Schia 2020].

Safe and Reliable AI has received a lot of attention in the European region recently compared to the US 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 lifecycle. 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. They enhance 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 [Barrett 2021]. If perceived as top-down oversight from experts, HITL is unlikely to address public trust deficits. Society-in-the-Loop (SITL) seeks broader consensus by extending HITL methods to larger demographics [Rahwan 2018] [Larson 2019], 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[9], where representatives of local stakeholder groups provide feedback during project lifecycles. SITL

[9] https://datapopalliance.org

combined with mass data literacy [Bhargava 2015] 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 [Nie 2020] [Kanchinadam 2020] using humans to label uncertain or subjective data points during training cycles. *Interactive sense making* [Middleton 2020] and *explainable AI* [Belle 2020] can also 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, US 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 US 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.

Defence - General Sir Patrick Sanders, head of UK Strategic Command, recently emphasized, "Even the best human operator cannot defend against multiple machines making thousands of manoeuvres per second at hypersonic speeds and orchestrated by AI across domains" [Ministry of Defence 2021]. 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 & security - Law enforcement is more nuanced. Though used only for warnings, Singapore's police robots have provoked revulsion in European press [The Guardian 2021], and the EU's AI Act reflects this attitude by classifying law enforcement as high risk. Some groups have claimed ambiguities in the EU's AI Act leave the door open for bias, unrestrained surveillance and other abuses [Skelton 2021], 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 into diagnostic accuracy shows that AI can improve healthcare outcomes [Prabhakar 2021] [Rangarajan 2021] [Bhandari 2020] [Gumbs 2021]. However, starting with patients and physicians, trust cascades upward, and, as Covid has shown, trust is ultimately political and thus needs to be carefully nurtured.

Transportation - Self-driving cars may receive the most publicity, but AI is also 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. Whom does the public trust to safely operate a vehicle?

Future Outlook

We think that 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 are currently exploring aspects of bringing humans into the loop for co-design and training of AI systems and human-machine teaming. We think that 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 and 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 guidelines [Ayling 2021] to frameworks for national or transnational 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.

Acknowledgements

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).

Authors

Stuart E. Middleton, Lecturer in Computer Science, University of Southampton, Southampton, UK

Emmanuel Letouzé, Marie Curie Fellow, Universitat Pompeu Fabra, Barcelona, Spain

Ai Hossaini, Senior Visiting Research Fellow, Kings College London, UK

Adriane Chapman, Professor in Computer Science, University of Southampton, Southampton, UK

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. Paterson, C. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. ACM Comput. Surv. 54, 5, Article 111 (2021), 39 pages. https://doi.org/10.1145/3453444

Ayling, J. Chapman, A. Putting AI ethics to work: are the tools fit for purpose?, AI Ethics (2021) Doi: 10.1007/s43681-021-00084-x

Barrett, P. Hendrix, J. Sims, G. How tech platforms fuel U.S. political polarization and what government can do about it, (2021) https://www.brookings.edu/blog/techtank/2021/09/27/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/

Belle, V. Papantonis, I. Principles and Practice of Explainable Machine Learning, arXiv, (2020) doi: arXiv:2009.11698

Bhandari, M. Zeffiro, T. Reddiboina. M. Artificial intelligence and robotic surgery: current perspective and future directions. Curr Opin Urol. (2020) 30(1):48-54. doi:10.1097/MOU.0000000000000692

Bhargava, R. Deahl, E. Letouzé, E. Noonan, A. Sangokoya, D. Shoup, N. Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data, Data-Pop Alliance White Paper (2015)

Bruynseels, K. Santoni de Sio, F. van den Hoven, J. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. Front. Genet. 9:31. (2018) doi:10.3389/fgene.2018.00031

DSTL, AI and data science: defence science and technology capability (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. Elyan, E. Artificial Intelligence Surgery: How Do We Get to Autonomous Actions in Surgery? Sensors (Basel). (2021) 21(16):5526 doi:10.3390/s21165526

Hartmann, K. Steup, C. Hacking the AI - the Next Generation of Hijacked Systems, 2020 12th International Conference on Cyber Conflict (CyCon), (2020), pp. 327-349, doi:10.23919/CyCon49761.2020.9131724.

Hunt, E.R. Hauert, S. A checklist for safe robot swarms, Nature Machine Intelligence (2020) doi: 10.1038/s42256-020-0213-2

Kanchinadam T. Westpfahl, K. You, Q. Fung, G. Rationale-based Human-in-the-Loop via Supervised Attention, 1st Workshop on Data Science with Human in the Loop, DaSH@KDD2020 (2020), August 24, 2020, Virtual Conference

Kania, E.B. Chinese Military Innovation in the AI Revolution, The RUSI Journal, 164:5-6, (2019) 26-34, DOI: 10.1080/03071847.2019.1693803

Kerasidou, C. Kerasidou, A. Buscher, M. Wilkinson, S. Before and beyond trust: reliance in medical AI, Journal of Medical Ethics (2021) doi: http://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

Middleton, S.E. Lavorgna, L. Neumann, G. 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 Defence, Commander of Strategic Command RUSI conference speech (2021) https://www.gov.uk/government/speeches/commander-of-strategic-command-rusi-conference-speech

Morton, S. Booth, M. The EU's "Third Way" to AI Regulation (2021), https://www.internetandtechnologylaw.com/eu-third-way-ai-regulation/

NHS-X, Cancer digital playbook (2021), https://www.nhsx.nhs.uk/key-tools-and-info/digital-playbooks/cancer-digital-playbook/

Nie, Y. Williams, A. Dinan, E. Bansal, M. Weston, J. 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. 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, Volume 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

Rangarajan, A.K. Ramachandran, H.K. A preliminary analysis of AI based smartphone application for diagnosis of COVID-19 using chest X-ray images, Expert Systems with Applications, Volume 183, (2021) doi:10.1016/j.eswa.2021.115401

Roberts, H. Cowls, J. Morley, J. Taddeo, M. Wang, V. 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

Rousseau, D.M. Sitkin, S.B. Burt, R.S. Camerer, C. Not So Different after All: A Cross-Discipline View of Trust, Academy of Management Review, 23, (1998) 393-404, doi: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 Association (2021) https://beyondstandards.ieee.org/the-ieee-certifaied-framework-for-ai-ethics-applied-to-the-city-of-vienna/

Schia, N.N. Gjesvik, L. Hacking democracy: managing influence campaigns and disinformation in the digital age, Journal of Cyber Policy, 5:3, (2020) 413-428 doi:10.1080/23738871.2020.1820060

Skelton, S.K. NGO Fair Trials calls on EU to ban predictive policing systems, ComputerWeekly (2021) https://www.computerweekly.com/news/252506851/NGO-Fair-Trials-calls-on-EU-to-ban-predictive-policing-systems

Taddeo, M. McCutcheon, T. 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

The Guardian, via Agence France Presse, 'Dystopian world': Singapore patrol robots stoke fears of surveillance state', (2021) https://www.theguardian.com/world/2021/oct/06/dystopian-world-singapore-patrol-robots-stoke-fears-of-surveillance-state

RUSI-TAS, Trusting Machines? Conference (2021) https://www.tas.ac.uk/eventslist/trusting-machines/trust-machines-conference-programme/

van der Valk, H. Haße, H. Möller, F. Arbter, M. Henning, J. Otto, B. A Taxonomy of Digital Twins, AMCIS 2020 Proceedings, (2020)

WeVerify, Horizon 2020 grant agreement 825297 (2021), https://weverify.eu