TRUST &
TECHNOLOGY
INITIATIVE

Digital Future(s)

2020 Research Perspectives

The Future of Digital Technologies and their Societal Impact

collated by the Trust & Technology Initiative

UNIVERSITY OF CAMBRIDGE

Table of Contents

Introduction: Digital Future(s)

2020 Research Perspectives
on the
Future of Digital Technologies and their Societal Impact

Around the world, people are both growing more dependent on, and simultaneously distrustful of, digital technologies. Privacy, veracity and safety concerns are rising, as is the fear of how automation and artificial intelligence will impact lives and livelihoods.

Yet digital technologies are likely to become even more prevalent - as well as more sophisticated - over the next decade, raising ever more fundamental questions for inventors, providers, regulators and users of these technologies, if the digital impact on society is to be beneficial. These are being thrown into even sharper relief by the COVID-19 pandemic and its accelerating effect on digitisation.

In response to the increasing pervasiveness of digital technologies a body of research on design and governance of trustworthy next generation technology has emerged in recent years. Technological, sociological, legal, economic, political and journalistic perspectives are converging on an exciting cross-disciplinary space in which vital questions of utility and morality are being debated.

At this crucial global juncture, the Trust & Technology Initiative is gathering a series of essays on the interplay between society and digital technologies towards an overview of cutting-edge thinking on this topic.

We are also taking the opportunity to reflect on how wider conversations on trust and digital technologies have evolved since the launch of the Initiative as a Strategic Research Initiative in 2018; to document current clusters of expertise at Cambridge University and to seed new conversations beyond the 2020 horizon.

About the Trust & Technology Initiative

Vision, People, Themes, Activities

Trust & Technology - Vision

The Trust & Technology Initiative aspires to inform trustworthy design and governance of next generation technology.

The Initiative promotes informed, critical, and engaging voices in the light of digital technology's increasing pervasiveness in societies. It is unique in combining cutting edge deep technology competence with social science and humanities expertise, which enables the dynamic exploration of emergent use cases and realistic future scenarios.

Trust & Technology - People

The Trust & Technology Initiative brings together actors from multiple disciplines for the purposes of knowledge exchange and collaboration.

The Initiative serves as meeting place for collaborators from various backgrounds, fostering constructive dialogue between academia, industry, civil society and policy makers. It acts as a gateway to the Cambridge research ecosystem for external partners interested in trust & technology. Our network spans social science, legal, technical, and humanities disciplines as well as policy makers and industry.

Trust & Technology - Themes

The Initiative aims to showcase the Cambridge University research ecosystem around questions of trust and technology, such as:

- Relationships and interplays between technology and society;
- Legal, ethical and political frameworks of trust and technology;
- The nature of trust *in* technology, and *through* technology;
- Technical foundations of more trustworthy computer systems.

Currently, the Initiative is particularly engaged in topics concerning societal implications inherent in AI, and of organisational dominance, power and responsibility in the digital space.

Trust & Technology - Activities

The Trust & Technology Initiative engages with researchers and their partners, and enables effective discussion and collaboration.

As an interdisciplinary networking forum, the Initiative helps new research ideas to emerge and enables prototyping and testing of ideas. Collaborative research and engagement across several disciplines takes the form of joint events, co-operation on specific research projects, as well as outreach activities and publications aimed at the general public.

Get involved

The Trust & Technology Initiative is open to all interested researchers at Cambridge University who share the Initiative's aims. We also welcome external collaboration inquiries.

There is a variety of ways to get involved more closely with the Initiative and help drive its research agenda, such as hosting events, talking on panels, chairing events, helping organise events, producing publications. The Initiative operates as a 'big tent' for volunteer initiative from different disciplines; activity is guided by the interests and projects of our membership.

If you would like to work with the Initiative in some way, please email admin@trusttech.cam.ac.uk, including a few sentences about your research and interests and how they relate to Trust & Technology, and whether you would like to play an active part in the Initiative.

Get in touch

- Website: www.trusttech.cam.ac.uk
- Twitter: @CamTrustTech
- Mailing list: bit.ly/CamTrustTechList
- Email: admin@trusttech.cam.ac.uk

2020 Research Perspectives

What is trust in technology? Conceptual bases, common pitfalls and the contribution of trust research

Dr Frens Kroeger
Centre for Trust, Peace and Social Relations, Coventry University

In a fundamental sense, all technology depends on trust. What makes technology 'technology' is precisely the fact that most users do not know – and do not need to know – how it works; instead, they hold the confident positive expectation that a mechanism which is ultimately opaque to them will bring about the desired outcome.

Consequently, when we talk about technology, we also need to talk about trust. After all, technologies can work flawlessly but still be rejected by an untrusting audience; conversely and potentially even worse, deeply flawed technologies can spread to all corners of the globe based on the misplaced trust of users.

While it is highly positive that the discussion on trust in technology is widening, being taken up by more and more technology experts, it still needs deepening. All too often in the newly emerging research on trust and technology there appears to be an implicit assumption that it is the specific technology under investigation that lends complexity and intrigue to the topic, whereas trust is presumed to be more or less self-explanatory. Often, this seems to be driven by the idea that after all, we all know how trust works in our daily practice; of course, if this were a valid hermeneutical principle, social science as a whole would be largely redundant.

Equally often, trust is tacitly equated with other terms (like security, confidentiality, or risk, to name but a few) which turn out to be the real focus of the work presented, with trust shoe-horned in as an afterthought that only seemingly links the piece to a novel debate. This is especially visible in the case of conference presentations, which will often mention trust in the title, the introduction and conclusion but nowhere in the main body of the presentation, which

instead deals with the concept that is really at the centre of the researcher's interest. Similarly, many empirically oriented studies go to great pains to operationalise it but at a closer look, what is being operationalised are often other, related phenomena (for instance adoption or use of a technology, even though we know that users can adopt a technology without fully trusting it, while others may trust but still choose not to adopt it for a myriad different reasons).

Why is this problematic? When trust is chosen as a label, but no real rigour is invested into its understanding and conceptualisation, we end up talking at cross-purposes. For instance, when I was a member of the team that compiled the first annotated bibliography on Trust in Artificial Intelligence for the Partnership on AI (2019), a key problem we encountered in categorising the corpus of texts was that the majority of papers were not communicating with each other in any meaningful way; much of the time they were effectively talking about virtually unrelated problems.

This is liable to keep the study of trust in technology from achieving coherence and to reduce the potential for both insight and impact of research on the topic. Even worse, if we purport to talk about trust but fail to do so with conceptual rigour, any intervention we design may miss the mark and may facilitate the development of factually untrustworthy technologies.

I want to argue that the way of achieving the coherence required is to draw on the insights and concepts provided by trust research, as an established and mature field of study. The systematic study of trust, arguably starting with some of the early 20th century classics, has long solidified into its own research field, with dedicated conferences, professorial appointments and research centres, and it is imperative for the debate on trust in technology to draw more strongly on the rich insights this interdisciplinary field of study has produced particularly over the last 25 years, incidentally with several volumes edited by Cambridge scholars leading the charge (Gambetta, 1988; Lane & Bachmann, 1998). (For a brief overview of what we commonly refer

to as "trust research", see for instance the table of contents in Bachmann & Zaheer, 2008.)

What would be some of the first and most basic lessons we can draw from this rich vein of research? Of course, within this very brief format I cannot provide an exhaustive list, but we may think at least of a few of the most basic and fundamental toeholds that matter here. As a very first step on an admittedly long way, we could make sure that at a minimum we always distinguish clearly between different groups of trustors, trusted objects and trust dimensions. The straightforward question to ask for this would be: who trusts what, and in what respect?

While this may seem trivial, at this stage it is anything but that. In my research on trust in autonomous vehicles (AV), I often encounter simplistic surveys investigating "what percentage of people in country X trust AV", though in reality, this may mean different things: for instance, do respondents trust the AV merely to keep its driver safe, but at a closer look we would find that they are not confident the privacy of their data will be preserved?

On reflection, we may also note that there are further stakeholder groups which matter, and that their trust requirements differ from each other (Pirson & Malhotra, 2011); for instance, other road users will want to trust that AV are safe not just for drivers but for cyclists, pedestrians and pets too; car rental companies may choose to focus on reliability and cost effectiveness; and insurance agencies need to be assured regarding the legal liabilities created by autonomous driving.

Even the question of what is being trusted may not always be as straightforward as it may seem at first. For instance, do the trust problems which are frequently diagnosed in regard to AI (Partnership on AI, 2019) relate to users' distrust of the algorithm as a technology, to the purposes for which the algorithm is being employed, or even to the organisation developing and deploying the algorithm?

(For algorithms making recommendations on consequential matters as different as bail, grade distributions or children's social care, it

makes a big difference whether suspicion relates to the data that the algorithm was trained on or to the question whether the respective agency intends to use technology as a pseudo-objective justification for a socio-political agenda.) To complicate things further, each of these different objects has identifiable analogues of the ability, benevolence and integrity that we look for in human trustees, and the relationships between the interlinked trust objects situated across different analytical levels are complex and non-trivial (Kroeger, 2012, 2017).

Embracing these and many more advanced concepts and mechanisms – from the genesis of System Trust over the preconditions for rapidly evolving Swift Trust to the possibility of simultaneous trust and distrust – and contextualising them to the unique setting of individual technologies will enable the study of trust in technology to make rapid advances as a coherent field whose research findings relate to each other in ways that enable fruitful communication and add value both to individual studies and to the field as a whole. First and most importantly, however, I think we will all need to agree on one thing: when we talk about trust and technology, we need to give both equal attention. Leveraging the insights that trust research has created over the last decades will be a central tool in this endeavour.

References

Bachmann, R. & Zaheer, A. (eds.) (2008). *Landmark Papers on Trust*. 2 vols. Cheltenham: Edward Elgar.

Gambetta, D. (ed.) (1988). *Trust: Making and Breaking Cooperative Relations*. Oxford: Blackwell.

Kroeger, F. (2012). Trusting Organizations: The Institutionalization of Trust in Interorganizational Relationships. *Organization* 19: 743-63.

Kroeger, F. (2017). Facework: Creating trust in systems, institutions and organisations. *Cambridge Journal of Economics* 41: 487-514.

Lane, C. & Bachmann, R. (eds.) (1998). *Trust Within and Between Organizations: Conceptual Issues and Empirical Applications*. Oxford: OUP.

Partnership on AI (2019). Human-AI Collaboration Trust Literature Review – Key Insights and Bibliography.

Available at https://www.partnershiponai.org/human-ai-collaboration-trust-literature-review-key-insights-and-bibliography

Pirson, M. & Malhotra, D. (2011). Foundations of Organizational Trust: What Matters to Different Stakeholders? *Organization Science* 22: 1087-104.

Dr Frens Kroeger is an Assistant Professor (Senior Lecturer Level) at the Centre for Trust, Peace and Social Relations at Coventry. He is an alumnus of Corpus Christi College, University of Cambridge, and has taught at universities in the UK, Germany, Switzerland and Japan. He has studied the phenomenon of trust across a wide variety of different contexts for over 15 years, and his research on the topic has been published in leading journals in the field.

Would you trust a cybercriminal? Exploring the cold start problem in an online cybercrime market

Dr Alice Hutchings

Cambridge Centre for Cybercrime, Department of Computer Science and Technology, University of Cambridge

Would you trust a stranger? Or an unknown business? How much trust would you place in a known cybercriminal? We often hear about the risk of cybercrime, but what is at stake for the crooks? Much cybercrime requires some level of reliance on others, whether it be renting infrastructure to run botnets, or cash-out services to monetise illicit gains. When an offender is cheated, they can't complain to the police, or take the other party to court. If it were a face-to-face interaction, criminals may at least be able to retaliate using violence if cheated, but online their opponents may be anonymous and protected by physical distance.

It is for these reasons I, along with my colleagues at the Cambridge Cybercrime Centre, am interested in trusting the untrustworthy within

cybercrime communities, where information asymmetry abounds. Trust tends to be developed over time, with repeated interactions. Criminals can also rely on signalling mechanisms developed to facilitate trust, such as reputation systems with underground markets, which are clearly modelled on eBay and Amazon recommendation systems.

The use of escrow services can also mitigate some of the risks (although perhaps displacing them to those that operate such services, who can run exit scams).

One of the great unknowns, however, is how new entrants to the cybercrime scene can establish this much-needed reputation. In economics, this is known as the *cold start problem*—the conundrum faced by new actors who find that others do not want to trade with them due to lack of reputation, but they cannot gain reputation as nobody will trade with them.

We recently had the opportunity to explore how the cold start problem may be overcome in an online black market. We collected data relating to 190,000 contracts from a new reputation system that had been set up in an established cybercrime market. The market traditionally provided a place for advertisements, but did not facilitate transactions. However, due to reports of scammers (often referred to as 'rippers' within cybercrime communities), a new market system was established. This includes logging contracts between users, which are then visible to those who pay a small fee. This new system provides users with a way to dispute transactions, and acts as a recommendation system to signal trustworthiness to potential buyers.

We explored this data, which spanned two years, over three discrete periods, which we called the set-up, stable, and COVID-19 eras. The first era, set-up, contained contracts made voluntarily on the market. The stable era starts when contracts became compulsory, while the COVID-19 era begins when the global pandemic was declared by the World Health Organisation. In our paper (Vu et al., 2020) we track the

effects of the pandemic on this cybercrime market, concluding that it stimulated, but did not transform, the market.

We found the most common marketplace activity was the provision of cash-out services, transferring currency from one type to another. The most exchanged currency types are Bitcoin and PayPal, and the funds exchanged are presumably obtained illicitly. We found most cold starters (new actors joining the market during the stable period) started to gain their reputation by engaging in low-level currency exchange, gradually increasing as they became more trusted on the market. In this way, the contract system allowed them to signal their experience and trustworthiness. Over time, including during the pandemic, we observed an increasing trend towards greater concentration of a few key actors on the market, who accounted for a disproportionately high number of transactions.

References

Vu, A. V., Hughes, J., Pete, I., Collier, B., Chua, Y. T., Shumailov, I., & Hutchings, A. (2020). *Turning up the dial: The evolution of a cybercrime market through set-up, stable, and COVID-19 eras.* Proceedings of the ACM Internet Measurement Conference, Pittsburgh.

Dr Alice Hutchings is a University Lecturer in the Security Group at the Computer Laboratory, University of Cambridge, a Fellow of King's College and Deputy-Director of the Cambridge Cybercrime Centre. Bridging the gap between criminology and computer science, Alice's research interests include understanding cybercrime offenders, cybercrime events, and the prevention and disruption of online crime. The Cambridge Cybercrime Centre takes a data-driven approach to improving the quantity and quality of cybercrime research among academics and develops robust identifiers and evidence of criminal behaviour with a view to crime prevention and mitigation.

Covid-19 and the growth of digitization: How a human virus has affected trust in the digital ecosystem

Dr Jennifer Daffron
Centre for Risk Studies, Cambridge Judge Business School

The cyber risk ecosystem is made up systemic and cascading risks with the potential for massive disruptions to nations, businesses, and individuals to occur from a single source. In 2017, that source was a computer virus called WannaCry. This particularly nasty ransomware infected millions of Windows systems in over 150 countries causing hundreds of millions of dollars in damages, pain, and suffering across the entire world. WannaCry infected systems without concern for who was hurt; the NHS alone lost £92 Million[1]. Such criminality is not unusual but has become more the rule than the exception with infamous predecessors such as Melissa and ILOVEYOU creating similar devastation.

What makes WannaCry different it that the headlines it created marked a point in cyber security history where businesses leaders began to heed the "not if but when" warnings more seriously; cyber-attacks rose to the top of corporate risk registers all over the world. Despite the rise, underlying trust in the overall digital environment wasn't seriously impacted. Even with the disruption to global supply chains, the threat to human life, and loss to revenue caused by WannaCry and its predecessors, the trust in digital systems didn't waver. Digital growth didn't flag but continued to grow even by those infected by WannaCry.

This trust seems strange especially when, beginning in 2019, the world experienced a virus that generated a distinctly negative trust in digital systems. Doubly strange because this time the virus wasn't in computers but in humans. Covid-19 seems to have irreversibly affected trust in the digital ecosystem. Steps taken to contain the

[1] https://www.infosecurity-magazine.com/news/wannacry-cost-nhs-92-million

spread of the virus put millions of people under duress by forcing them to rely on less stable digital systems resulting in an explosion of growth in the digital economy—and a concomitant decrease in trust for digital systems.

The speed of the massive migration to work from home caught employees and employers off guard. Ad-hoc set-ups and personal devices in bedrooms and on kitchen tables replaced secure in-office digital environments where network access had been tightly controlled and with up to date security. This drastic mismatch instigated the first key area of increased digital dependence and accompanying insecurity. To combat the increased risk, companies moved to third-party monitoring capabilities, migrated work environments to the cloud, and invested in video conferencing software. The growth statistics of companies like Zoom and Microsoft's cloud services make the explosion of dependence undeniable. Zoom's daily user count increased by almost 3000% in just a little over one quarter with nearly 100 million users a day.[2] Microsoft officials say the company has seen a 775% increase in demand for its cloud services in regions enforcing social distancing.[3]

Connected devices and services also moved the Highstreet into the home, marking a second area of increased dependence. According to an Adobe report, total online spending in May 2020 hit $82.5 billion, up 77% year-over-year. In the UK, ecommerce took two decades to go from zero to around 7 per cent of total grocery sales. It then went from 7%to 13% in about eight weeks as Highstreet shut down.[4]

---

[2] https://medium.com/swlh/zooming-ahead-the-explosive-growth-of-zoom-during-the-pandemic-34f55b1f13e8

[3] https://www.zdnet.com/article/microsoft-cloud-services-demand-up-775-percent-prioritization-rules-in-place-due-to-covid-19/

[4] https://www.forbes.com/sites/kaleighmoore/2020/06/24/as-online-sales-grow-during-covid-19-retailers-like-montce-swim-adapt-and-find-success/#51f0d5886d78

It is difficult to say whether these trends will remain once the population can return to stores in person but for the foreseeable future forced dependence is undeniable.

Like all explosive reactions, the sudden and accelerated rate of change created instability. Recurring failures in the security of home offices and the unreliability of the jury-rigged technologies and services that scramble to support them reduces trust in the digital ecosystem as a whole. The millions of dollars being spent on cyber security for home offices was no match for the dark side of social engineering that entered into inboxes under the guise of the world health organisation, PPE providers, and job retention schemes. The opportunistic and morally corrupt nature of cyber criminals exploited the fears and uncertainty of individuals as the news spread of illegal access to personal and professional networks around the world. Scams increased by 400% over the month of March, making Covid-19 the largest ever security threat.[5]

Misinformation and fake news campaigns about the virus furthered the distrust in digital technologies and services at a time when accurate information was more important than ever. Rumour mongering about the virus including its scale, prevention, treatment and source circulated through all online mediums followed closely by denials that seemed as crazy as original tale. A Rutgers-led[6] study found that online misinformation, or "fake news," lowers people's trust in mainstream media.

This distrust escalated to conspiracy theories against technologies—both digital and medical. In February 2020, BBC News reported that conspiracy theorists on social media groups found an alleged a link between coronavirus and 5G mobile networks, claiming that both the Wuhan and Diamond Princess outbreaks were directly caused by the

---

[5] https://www.reedsmith.com/en/perspectives/2020/03/coronavirus-is-now-possibly-the-largest-ever-security-threat

[6] https://phys.org/news/2020-06-fake-news-lowers-mainstream-media.html

introduction of 5G and wireless technologies.[7] Scientists have highlighted the complex ties between the virus and misinformation and warned that developing a working Covid-19 vaccine "might not be enough" to end the pandemic unless steps are taken by governments and technology firms to tackle coronavirus misinformation.[8]

Correlations between different types of risk are not new. To create accurate correlations, one must understand the probability and likelihood that these events can occur separately and the key variables of where they are linked. A computer virus does not always affect global cyber security. A human virus does not always affected global health security. But as our world grows digitally and we enter into the fourth industrial revolution it is essential that these correlations are revisited. It is even more essential that the human variable of "trust" is entered into the equation. A human virus does not always affect global cyber security, but this did. It is yet to be seen if the digital economy can withstand the weight of the dependence we have put upon it and whether or not our trust was misplaced.

Dr Jennifer Daffron is a Research associate at the Cambridge Centre for Risk Studies, where leads the research on digital risk. Her work defines and exposes cyber threat vulnerabilities on organisational and human behavioural platforms for companies around the world. Jennifer holds a PhD in Experimental Psychology from the University of Cambridge and has published several papers on attentional templates in visual search.

- Website: www.trusttech.cam.ac.uk
- Twitter: @CamTrustTech

[7] https://www.bbc.co.uk/news/technology-51646309

[8] https://www.sundaypost.com/fp/covid-19-misinformation-vaccine-alone-might-not-be-enough-to-end-pandemic/

Reflections on the 'Critical Perspectives on Law, Technology, & Society' Reading Group

Lily Hands; PhD Candidate
Faculty of Law, University of Cambridge

In 'The New Forms of Control', Herbert Marcuse opined that a 'comfortable, smooth, reasonable, democratic un-freedom prevails in advanced industrial civilization, a token of technical progress.'[1] If so, what is the nature of this technological 'unfreedom', and what (if anything) does it have to do with law?

This question animated both my doctoral research, and my participation in the Critical Perspectives on Law, Society and Technology Reading Group. Convened by Drs Jennifer Cobbe and Christopher Markou, the reading group set out to explore how technological change interacts, challenges, subverts, and co-evolves with the law and society. The group brought together a diverse spectrum of people from disciplines including law, computer science, sociology and economics, among others. The 'productive friction' engendered by these disparate viewpoints – expertly steered by Jenn – led to several breakthroughs in my work, both major and minor.

On a general level, 'zooming out' from law helped me establish a broader perspective on its role and limitations, something which can be difficult to achieve from within the confines of one's own discipline. The interdisciplinarity of my dissertation has unquestionably been strengthened by watching different modes of knowledge cross-pollinate in real time. It has also provided an excellent forum to meet others who share some of my interests and concerns, and to learn more about their unique perspectives.

On a substantive level, the radical consequences of emerging technologies for society and for law was made clear to me early on by Marshall McLuhan's *Understanding Media: The Extensions of*

[1] Herbert Marcuse, *One Dimensional Man* (Beacon Press 1964), 34.

Man, which the group discussed in one of its initial sessions. McLuhan argues that the medium of a communication shapes and controls the form, scale, pace and pattern of human relations and action, regardless of its use or content. Identifying patterns within a medium as well as in its relationships with other media can reveal, predict and possibly control the social and normative effects of their use. Conversely, without a detached focus on the nature of specific media, the assumptions, biases and values they inherently afford – and thus their functional limits for humanity – will remain invisible.[2]

These reflections were influential in the development of my first year PhD report, which focused in part on the shift in media from natural language to binary code and its implications for law as a system based on an underlying concept of justice.

Beyond McLuhan's arguments, the group enabled me to better appreciate how the relationships between law, society and technology are fundamentally co-evolutionary. Law and society do not just respond to technology; they shape the conditions of possibility of technology itself, including whether and how it comes into existence, its use, and the modes and discourses of acceptance and resistance. From this co-evolutionary approach, it also follows that there is no such thing as 'optimisation' within the systems we have created – whether social, legal or technological.

At the same time, by better understanding these dynamics and how they came to be, we can increase our potential range of action – the conditions of possibility under which we might more consciously decide how we would like to both use, and respond to the prospect of, emerging technologies. For example, one of the most enduring themes of discussion within the group was how powerfully – and perhaps pathologically – technology, law and society appear to be yoked to the economic system, even as emerging technologies

[2] Marshall McLuhan, *Understanding Media: The Extensions of Man* (MIT Press 1994) ch1.

render the logic of accumulation *within* that economic system more extractive and antidemocratic than under earlier forms of capitalism.[3]

Impersonal and rule-centred legal institutions including (but not limited to) property and contract have in many ways enabled this transition, by tending towards compatibility with a rationality of efficient administration. At the same time, law itself is being transformed by new, often algorithmic, modes of organisational efficiency.[4] This might lead us to ask just how close we are to the gloomy totalitarian vision of Vaclav Havel in *The Power of the Powerless*:

'Technology—that child of modern science, which in turn is a child of modern metaphysics—is out of humanity's control, has ceased to serve us, has enslaved us and compelled us to participate in the preparation of our own destruction. And humanity can find no way out: we have no idea and no faith, and even less do we have a political conception to help us bring things back under human control.

And so we return to my original question: how can, and should, we respond to the spectre of 'unfreedom' identified by Havel and Marcuse decades ago, and which more recent analyses of the contemporary technological, legal and social landscape have only tended to reinforce? And how can we respond without needlessly jettisoning legitimate and useful technologies and their uses?

I am yet to answer these questions. The intellectual starting point of my PhD was Karl Polanyi, who argued that freedom can only be achieved *through* and *within* society – individuals must have the means and opportunity to take full responsibility for all the con-

[3] See eg Shoshanna Zuboff, 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilisation' (2015) 30 Journal of Information Technology 75.

[4] David Lyon, 'Surveillance Society' (Festival del Diritto, Piacenza, Italy, 28 Sep 2008) http://www.festivaldeldiritto.it/2008/pdf/interventi/david_lyon.pdf accessed 03/02/2020.

sequences of their actions.[5] Using Polanyi's social freedom as a compass, I am currently researching the nature of the risks posed by artificial intelligence and their implications for the legal system's capacity to respond and co-evolve with such technologies in the future. The intellectual structure and interdisciplinary grounding I received from the group continues to be a valuable resource in this respect.

Lily Hands is a PhD candidate in Law at the University of Cambridge. She worked as a judge's associate and litigator before graduating with a Master of Law from Cambridge in 2018. Her current research focuses on the relationship between law, risk and AI.

Why the AI impacts ecosystem must move beyond 'near-term' and 'long-term'

Dr Jess Whittlestone and Dr Shahar Avin
Centre for the Study of Existential Risk, University of Cambridge

The impacts of AI are already visible in numerous domains, while research breakthroughs are likely to precipitate even greater impacts than those we are already seeing. The combination of algorithmic bias, increasing technological unemployment, and AI concentrating power in the hands of tech companies could entrench existing patterns of systemic discrimination and lock in much more extreme global inequality than we see today.[1] [2]

[5] Karl Polanyi, The Great Transformation: The Political and Economic Origins of Our Time (2nd ed., Beacon Press 2001).

[1] West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute, 1-33.

[2] Lee, K. F. (2018). AI superpowers: China, Silicon Valley, and the new world order. Houghton Mifflin Harcourt.

If advanced language models become a regular component of fake online personas, this could corrupt our information ecosystem to the extent that "the pillars of modern democratic self-government—logic, truth, and reality—are shattered".[3] [4] The increasing integration of machine learning systems in critical infrastructure across the world holds huge promise for improving critical resource management but could also open up huge vulnerabilities, where accidents could result in huge loss of human life. More generally, AI technologies might put pressure on international law by driving frequent changes in diverse sectors, putting stress on existing treaty regimes and inhibiting effective global governance.[5]

As many have pointed out, research on the impacts, risks and governance of AI so far has tended to 'cluster' into two groups one focused on identifying and shaping the impacts of existing and imminent applications of AI in society ('near-term'), and the other focused on the potential existential risks of developing human-level AI ('long-term'). However, as the research community has made progress on both near- and long-term issues, we are beginning to see the limitations of both these approaches.

While immediate issues resulting from current applications of AI are things we can address now, and may still be very important to address in the long-run, this 'near-term' focus must inevitably be somewhat reactive to problems as they become apparent. For example, wide-spread acknowledgement that algorithmic bias and data privacy are

---

[3] Lin, H. (2019). The existential threat from cyber-enabled information warfare. Bulletin of the Atomic Scientists, 75(4), 187-196.

[4] Seger, E., Avin, S., Pearson, G., Briers, M., Ó hÉigeartaigh, S.S., and Bacon, H. (2020). Tackling threats to informed decision- making in democratic societies: Promoting epistemic security in a technologically-advanced world. Alan Turing Institute.

[5] Maas, M. M. (2019). International law does not compute: Artificial intelligence and the development, displacement or destruction of the global legal order. Melb. J. Int'l L., 20, 29.

serious ethical problems has come in response to highly-publicised mistakes including racial bias in parole rating algorithms[6] and data breaches as a result of collaborations between DeepMind and the Royal Free Hospital in the UK.[7] As AI systems become more sophisticated and integrated into more important areas of society, the stakes of 'mistakes' will only get bigger, and addressing problems after-the-fact becomes increasingly infeasible.[8]

On the other hand, while low-probability, extreme-stakes risks from human-level AI are worth preparing for, the abstract nature of these concerns and broad assumptions involved make it difficult to know how these concerns should guide decisions about the development, deployment, and governance of AI today. This is an instance of the 'Collingridge Dilemma': before a technology is well-developed it is difficult to predict its impacts, but once those impacts are more apparent it is often too late to change them. By exploring the possible applications and impacts of current research trends in AI over the next 5-15 years, we may be able to find a 'sweet spot' where impacts are grounded enough in current trends to prepare for now, but far enough in the future to not already be entrenched.

The ecosystem addressing AI's impacts on society must diversify beyond the purely 'near-term' or 'long-term' (though this doesn't mean every group or sub-community must do so). To ensure that work today on AI impacts, risks, and governance stays relevant and useful as capabilities advance, we must look ahead to consider possible emerging applications and impacts of AI, and identify actions we can take today that are likely to be robustly beneficial and mitigate risks

[6] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. Pro Publica.

[7] Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. Health and technology, 7(4), 351-367.

[8] Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. Research Policy, 42(9), 1568-1580.

across a range of scenarios. To ensure that we're able to prepare for and mitigate these most extreme impacts, we must more thoroughly explore different possible trajectories of AI development, deployment, and impacts, rather than centering all attention on preparing for a subset of scenarios in which human-level AI arises suddenly. As well as identifying areas of risk and future concern, there is also an urgent need to build shared visions of the future we want to create with AI, which can guide the development and use of this technology today.[9]

Exploring these 'mid-term' issues will require thinking rigorously about new methodological approaches. To anticipate and prepare for future impacts of AI, we must draw on the perspectives of a wide range of stakeholder groups, to bring domain expertise and ensure consideration of a diverse range of concerns. In addition, unlike short-term AI impacts, exploration of the "medium-term" requires more direct and prolonged engagement from the AI research community to identify plausible technology futures. Ensuring that future scenarios are grounded in an understanding of technical capabilities is particularly important given that our intuitions are often poor guides for the behaviours of future intelligent systems. While a broad range of tools and methods are available for AI futures exploration,[10] existing approaches tend to prioritise either deep expertise or diverse participation: none are perfectly suited for combining the two. We must therefore find novel ways to combine existing methods to bring deep technical expertise and diverse stakeholder groups together.

---

[9] Ramos, J., Sweeney, J. A., Peach, K. and Smith, L. (2020). Our futures: by the people, for the people. How mass involvement in shaping the future can solve complex problems. Retrieved from https://media.nesta.org.uk/documents/Our_futures_by_the_people_for_the_people_WEB_v5.pdf

[10] Avin, S. (2019). Exploring artificial intelligence futures. Journal of AI Humanities. Available at https://doi. org/10.17863/CAM, 35812.

Humans are not mere bystanders in this "AI revolution":[11] the futures we occupy will be futures of our own making, driven by the actions of and interactions between technology developers, policymakers, diverse stakeholders and numerous publics. There is therefore an urgent need to develop "anticipatory" approaches to the study of responsible AI.

Dr Jess Whittlestone is a member of the Trust & Technology Initiative's 2018-2020 Steering Group. For her biography, please see p X.

Dr Shavar Avin is a Senior Research Associate at the Centre for the Study of existential Risk. His research examines challenges and opportunities in the implementation of risk mitigation strategies, particularly in areas involving high uncertainty and heterogeneous or conflicting interests and incentives. Mixing anthropological methods and agent-based modelling, Shahar works to identify and design opportunities for impact.

Analyzing citizen data practices and how they challenge or work within dominant data regimes

Prof Jennifer Gabrys,
Department of Sociology, University of Cambridge

The evolving relationship between citizens and data is a fundamental issue of our time. It impacts social formation, cohesion and civil rights, since data has become the basis for innumerable social, political and economic processes and decisions. While data can contribute to original social insights, at the same time numerous concerns have arisen, ranging from the pervasive tracking and surveillance, to ownership monopolies that restrict access and control for data analysis, and production. In order to address these concerns, people

[11] Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90, 46-60.

are engaging in alternative practices of production, ownership and data analysis. Through these practices they are attempting to challenge dominant data regimes by becoming active in the creation of alternative practices and infrastructures. New data democracies are emerging. The Citizen Data project seeks to understand them in order to identify changing formations of citizenship, and to build more effective relations to data.

Through the CamPo funding initiative, this pilot research project investigates citizen data practices as constituting a crucial movement toward greater public participation in social, technological, political issues. By attending to the social and cultural aspects of citizen data and environmental data we investigate how data is collected, to whom data is valuable, how citizens challenge data regimes, and how this informs practices of citizenship. By working across theoretical, practical and policy-oriented engagements, the project analyzes how citizens continuously produce data through their digital exchanges, as well as how citizens are changing existing dynamics by generating their own data by producing new public goods and informational commons.

Among the wide array of issues impacted by citizen data practices, the Citizen Data project will focus on citizen-generated environmental data. The project will focus on the mode of production and the consequent actions that those data might entail. By rigorously attend to the social and cultural aspects of environmental data it will ask: how is data collected, to whom is our data valuable, how do citizens challenge data regimes, and how does this inform our roles as citizens? By working across theoretical, practical and policy-oriented engagements, the project will investigate how citizens continuously produce data through their digital exchanges, as well as how citizens are changing existing dynamics by generating their own environmental data producing new public goods and informational commons.

Through engaging with these emerging practices, we will be guided by three key research questions:

1. How are environmental data formed and generated *about* citizens? Whether through online tracking of energy use, the combining of environmental big data sets by technology companies, or the monitoring of environmental behaviours, citizens are the focal point for data collection and synthesis. We will focus on these dynamics, technologies and actors to analyse and document the consequences of citizens as objects of ongoing data collection and management.
2. How are environmental data collected and contributed *by* citizens? Citizens are now active in generating new datasets that are mobilised as alternative forms of evidence. We will investigate how movements such as citizen science, crowd sourcing, co-creation and open data are changing the processes whereby expertise is constituted and evidence is formed, circulated and acted upon.
3. How are environmental data dynamics constitutive *of* citizens and citizenship? New responsibilities are forming through citizens' claims to data rights, like access, review and amend data about *them*. We will examine how people become political subjects by making demands about and engaging with data in relation to environmental problems. In this way, we will consider how both citizens and citizenship are not only implemented, but also come into being through new relations to data.

In order to address these foundational questions, we will attend to the ways in which people become citizens through the things they do and the claims they make with environmental data. Rather than attempt to define the citizen in advance as a fixed category, we will instead scrutinise how the shifting practices and politics of citizen data generate new significant public formations for the environment as the conjunction of political, democratic and technological issue.

The research will take a project-oriented approach in two interrelated project areas: **Environmental Sensing** and **Climate Change**. Rather than merely reflect upon data-related policies and projects, we will create a *laboratory for citizen data* that will be activated through practical workshops. In these workshops, we will engage in pilot

creative and critical activities and experiments by working with a diverse range of actors and collaborators. In order to achieve transformational contributions to an emerging area of concern, we will address three key objectives:

1. Assemble an international community of transdisciplinary researchers, practitioners, and technologists within the space of project workshops and events to critically interrogate and practically investigate citizen data in order to foster more democratic practices.
2. Develop and exchange citizen data knowledge, techniques and practices through key project areas of data influence related to environmental sensing and climate change, in order to transform existing structures of expertise and digital exclusion.
3. Examine and create new infrastructures for citizen data that enable wider engagements with issues of environmental and digital participation, from the creation of public platforms and citizen archives to alternative social networks and data practices.

Through pursuing these objectives, the project will undertake a pilot-scale investigation into new formations of citizen data, with the aim to develop a new research collaboration across the Department of Sociology at the University of Cambridge and the médialab at Sciences Po. Working at the intersection of humanities and social science approaches, the research project will be informed by the broader disciplines of sociology, science and technology studies, design, digital studies, and environmental studies. By specifically investigating environmental data as a problem of technology and citizenship, this pilot project will seek to obtain follow-on funding in order transform data-focused scholarship in the social sciences and humanities.

Professor Jennifer Gabrys is a member of the Trust & Technology Initiative's 2018-2020 Steering Group; for her biography, please see p X.

A new model for oversight of technology companies?

Working with DeepMind Health: a review

Dr Julian Huppert
Intellectual Forum Jesus College, Cambridge

People are becoming far more aware of the dangers that can be caused by overwhelming power from the big technology companies. The Cambridge Analytica scandal and much else has highlighted the huge societal harms that can be caused.

I think the case for stronger legislation is very strong. Any overly powerful organisation can cause immense harm, either deliberately, or inadvertently – and even if you are sure that an organisation's current leadership is benevolent, how sure can you be that that will continue for the future?

But stronger legislation can only take us so far. It is a blunt instrument, no matter how hard you try to tweak it, it is almost impossible to eliminate bad outcomes, without preventing good ones or generating other problems. Indeed, the more arcane and byzantine laws get, the easier it can be for large organisations to find ways to game the system.

There is also a problem that legislation can only ever set a minimum standard. I would like to see a reason for companies to aspire to do more than the legal minimum, whatever that may be. For that reason I am very committed to driving enhancements in technology ethics, not as a way of avoiding legal regulation, but as a way of driving above and beyond mere legal compliance.

One specific example of this has been the work I did with an organisation called DeepMind Health (DMH). DeepMind, now owned by Alphabet, who own Google, is possibly the world's leading deep Artificial Intelligence company. They knew that when they went into healthcare, this would attract a lot of attention and criticism – aside from the normal sensitivities around health information, the idea that Google could get even more data quite rightly concerned many people.

As a result, they decided to try a new ambitious approach to oversight and governance, bringing in a panel of Independent Reviewers to keep an eye on them, and act as a sort of watchdog, giving advice and drawing public attention to any concerns and failings. I and eight others, all with some public prominence, were brought in as reviewers, and I was asked to Chair the group.

A core underpinning idea was that if you want to have trust, the best way to do that is to demonstrate trustworthiness. As a result, rather than relying on press coverage to argue you should be trusted, you find ways to be appropriately open and transparent, and hence demonstrate why you should be trusted – that should then lead to the trust deserved. Or, if trust isn't deserved, that will also be highlighted.

There were a number of features of this process that go well beyond the often-seen advisory groups, and mean that it was more than just ethics-washing. We were under no confidentiality requirements, but had access to any information we wanted (other than confidential patient data, for obvious reasons). We were explicitly free to share anything we wanted to share with the press and public if we felt that was appropriate. Indeed, our only real obligation was to produce an annual report in public – and DMH had no say on what we wrote. Additionally, we had completely free rein in what we chose to look at – nothing was off limits, and we had a budget to commission our own work, in whatever we felt was worth investigating.

Two examples from our first year perhaps illustrate how remarkable our freedom was. DMH were using an app called Streams to help clinicians at the NHS Royal Free Hospital identify acute kidney injury faster, potentially saving many lives. We wanted to see how securely the data was held, and how secure the app was, the coding environment, and anything else. We therefore commissioned an external security firm to go through everything from the code to the physical security of the data centre, and we then published their report – in full, including identifying the handful of failings that were noticed, none of which were serious.

How many companies would agree to have their code audited in this way, with the results published openly? This is normally unique to open source projects. I did ask Microsoft Health if they would consider this, and they said that they didn't need to, because they knew it was secure. I know if I was commissioning a major piece of software, I'd trust the people who openly admitted to some minor failings over those who asserted without proof that they had none.

Another example of our freedom was in regard to the legal position of the data sharing agreement that DMH had with the Royal Free Hospital. There were complaints that, among other things, DMH were operating beyond the role of a Data Processor, and had more control over the data that was legal. This led to an investigation by the Information Commissioner's Office, that lasted over a year.

Meanwhile, we commissioned our own independent legal advice, paid for by DMH, although I don't think they knew whom we commissioned until afterwards. That advice concluded that DMH had not broken the law – a view later reached by the ICO and others, who noted serious failings at the Royal Free. We would have published their conclusions whatever they had said – again few companies would voluntarily take that risk.

Our work developed in many other areas, such as looking at the clinical evidence base, the nature of public and patient involvement, both of which transformed as a result. We also set out a **set of 12 ethical principles** that we felt ought to apply to any technology company working in healthcare – and many would apply much wider.

[See opposite page]

After we had been going for 2 and a half years, our work was brought to an end by the end of DeepMind Health itself; a reorganisation meant that the research part of its work reverted to core DeepMind; and the applied part became part of Google Health. Neither group has used an equivalent approach.

1	2	3
Benefit to data providers The company seeks to ensure that patients, service users, healthcare systems and organisations, who are the source of data, benefit appropriately from the learning derived from it.	**Public, patient and practitioner engagement** The company proactively engages with patients, carers, practitioners and members of the public, and in response to their inputs	**Design for safety and utility** The company always designs products and processes that make it easier for staff to do the right thing and minimise unintended consequences.
4 **Evidence-driven** The company is committed to generating and sharing evidence of effectiveness for any interventions, including peer reviews as appropriate, and shall avoid over-claiming the effectiveness of any products and services.	**5** **Anti-monopoly** The company seeks to ensure that it promotes competition, and encourages other organisations, including SMEs, into the market: in particular, the company will ensure that their systems are interoperable, using open APIs.	**6** **A model employer** The company ensure that is exemplary in employment practices, including promoting diversity in all dimensions, equal pay, flexible working, and paying the living wage.
7 **Legal and ethical** The company obeys the letter and the spirit of all appropriate legislation and regulation, including taxation.	**8** **Protecting privacy** The company takes string steps to protect patients' privacy by design and in implementation.	**9** **Secure** The company continuously ensures the highest level of security of all data it holds.
10 **Transparency** The company promotes transparency in its own work and contracts, within the constraints of privacy.	**11** **Reasonable profit** The company will not use its assets or position to seek to extract excessive profits in its dealings with the public sector and will, as far as possible, operate contracts on an open book basis.	**12** **Openness** The company promotes a culture and maintains processes to encourage any member of staff to feel they can raise - without fear of adverse personal consequences - concerns they have about risks or unethical behaviour.

Did we succeed? I think it was a mixed bag. We definitely caused a number of improvements in the way DeepMind Health operated, and some of that has carried on in its new incarnations. We didn't succeed in demonstrating trustworthiness, though I think we did go somewhere along that line. One problem is the press – many similar approaches are just ethics washing, and many independent reports are far from that, being sanitised before release. As a result, they sometimes over emphasised any criticisms we did make, seeing them as the hints of a bigger iceberg underneath, whereas we kept to a warts-and-all approach.

I also think we didn't have long enough, nor sufficient profile for people to get used to looking at our work. There are also challenges around our own structure – why should people have had trust in us as appropriate proxies? We also discovered the limits of having had no confidentiality clauses – it meant that while we had the right to be told things in DMH, we couldn't be told things happening at the Alphabet level or with some others, where NDAs were needed for other reasons.

Overall, I think it was an excellent experiment. Like many experiments, it had successes and failings – and points out how to improve this approach next time.

Dr Julian Huppert is a member of the 2018-2020 Steering Group of the Trust & Technology Initiative; for his biography, please see page X.

At crucial junctures, the Trust & Technology Initiative gathers perspectives on the interplay between society and digital technologies towards an overview of cutting-edge thinking on this topic. Previous research perspectives are documented on our website: https://www.trusttech.cam.ac.uk/perspectives

Digital Security by Design: Toward computer systems that are more trustworthy

Prof Simon Moore
Department of Computer Science and Technology

All too often we see news of yet another attack on computer systems. In Cambridge we have been exploring how to redesign computer systems from the hardware up with the objective of making them fundamentally more trustworthy. After over ten years of research and over 150 researcher years of effort, we have produced the CHERI secure computer architecture[1] comprising new mechanisms for processors that provide the fundamental building blocks on which secure software can be built. We have prototyped complete systems to demonstrate the benefits of such an approach and are now in a process of transitioning the technology. Our aims are, out of necessity, ambitious: to change the entire computer industry to use our more secure technology. This is not something that can be done by spinning out a start-up company but requires engagement across the industry.

Through working with Innovate UK, the Digital Security by Design (DSbD) Industry Strategy Challenge Fund[2] was established in 2019 to transition the CHERI security technology. This comprises £70m of UK government funding and £117m of industry backing. When the funding was announced, Business Secretary Andrea Leadsom said:

- *"Cyber-attacks can have a particularly nasty impact on businesses, from costing them thousands of pounds in essential revenue to reputational harm.*
- *Cyber-criminals operate in the shadows, with the severity, scale and complexity of breaches constantly evolving. It's critical that we are ahead of the game and developing new technologies and methods to confront future threats, supporting our*

[1] https://www.cl.cam.ac.uk/research/security/ctsrd/cheri/

[2] https://www.cl.cam.ac.uk/research/security/ctsrd/cheri/dsbd.html

businesses and giving them peace of mind to deliver their products and services safely.

- *Investing in our world-leading researchers and businesses to develop better defence systems makes good business and security sense."*

Under the DSbD initiative, ARM is building the Morello platform that will demonstrate CHERI security on the ARM processor[3]. ARM Ltd has its HQ in the UK and is the world leader in processors for mobile phones, tablet computers, the Raspberry Pi, etc., and has recently been adopted by Apple for their new laptops and Apple mini using their M1 chip. The Morello platform hardware and software will be provided to academic and industrial partners to evaluate this new security technology and explore the myriad of software use-cases.

Microsoft's Security Response Center (MSRC) have already undertaken an analysis. Matt Miller led this work and in his talk at Bluehat 2019 he concluded that CHERI would have mitigated over 70% of all the vulnerabilities in Microsoft software in the last ten years. Such vulnerabilities include WannaCry that had a devastating effect on the NHS in 2018.

The DSbD initiative is also providing £10m of funding for nine UK research projects[4]. At the launch the Digital Secretary, the Rt Hon Oliver Dowden said:

> *"We have a world-class cyber security sector, and together we are working hard to make sure the UK is the safest place to work, connect and live online. With government support, these projects will build cutting-edge, secure technologies that will give people and businesses further confidence in our digital services and help weaken the threat of cyber attackers."*

[3] https://www.arm.com/blogs/blueprint/digital-security-by-design

[4] https://www.enterprisetimes.co.uk/2020/06/15/digital-security-by-design-nets-researchers-10-million/

To explore the social impact of these technologies, the DSbD initiative is funding the Describe research hub[5] hosted by University of Bath.

With all of the industrial and academic activity around CHERI, we have high hopes that this technology can be deployed into consumer produce and that in the longer term it will have a major impact, making computer systems more secure and robust, giving us a computer platform that is far more trustworthy than systems today.

Prof Simon Moore is the co-director of the Trust & Technology Initiative. For his biography, please see page X.

Embracing uncertainty: towards an innovative architecture for sensor-driven computing

Dr Phillip Stanley-Marbell
Department of Engineering

Existing computing systems largely treat sensor measurements as though they were error-free. As a result, computing systems that consume sensor data and which implement algorithms such as obstacle avoidance may perform billions of calculations per second on values that might be far removed from the quantities they are supposed to represent. When these algorithms control safety-critical systems, unquantified measurement uncertainty can inadvertently lead to failure of subsystems such as object detection or collision avoidance. This can in turn lead to injury or fatalities and the prospect and evidence of such failures reduces trust in autonomous systems.

Figure 1 shows one concrete example of measurement uncertainty in the individual points in a pointcloud generated by a LIDAR sensor:

[5] https://www.discribehub.org/

[… for separate download ..]

With the ever more pervasive use of sensors to drive computation and actuation such as in autonomous vehicles and robots which interact with humans, there is a growing need for computing hardware and systems software that can track information about uncertainty or noise throughout the signal processing chain.

The Physical Computation Laboratory research group is investigating new methods for quantifying how this uncertainty could affect algorithms which consume sensor data, as well as new classes of efficient algorithms (and hardware implementations) for making the best use of such sensor-level uncertainty characterization, with applications ranging from trustworthy autonomous systems to noisy intermediate-scale quantum computing.[6]

We investigate new ways to exploit information about the physical world to make computing systems that interact with nature more efficient and more trustworthy, by taking into account the noise and uncertainty inherent in all measurement in sensor-driven systems. Our research applies this idea to new hardware architectures for processing noisy or uncertain data, new methods for learning models from physical sensor data, and new methods for synthesizing state estimators (e.g., Kalman filters) and sensor fusion algorithms from physical system descriptions.[7]

Realizing these goals in practice involves equal measures of mathematical methods and theoretical work combined with the design and implementation of domain-specific languages and using the

[6] P. Stanley-Marbell, University of Cambridge (PI). EPSRC EP/V047507/1. Architectures and Distribution Arithmetic for Coupling Classical Computers to Noisy Intermediate-Scale Quantum Computers. January 2021 to January 2023.

[7] V. Tsoutsouras, S. Willis, and P. Stanley-Marbell. "Deriving Equations from Sensor Data Using Dimensional Function Synthesis". To appear, *Communications of the ACM,* Research Highlight, January 2021.

compilers for these domain-specific languages as the testbed for implementing new algorithms.

To ground our ideas, we frequently design processor- and FPGA-based custom hardware platforms as end-to-end testbeds.

Dr Phillip Stanley-Marbell is a member of the Trust & Technology Initiative's 2018-2020 Steering Group. For his biography, please see p X.

All contributions to the 2020 Research Perspective Booklet can also be accessed online at www.trusttech.cam.ac.uk.

The contents of the 2018-2020 Booklets and their respective download links are listed at:

https://www.trusttech.cam.ac.uk/perspectives/booklet-downloads

Handmaidens of Disinformation

Dr Rory Finnin
Department of Slavonic Studies

In 'La biblioteca de Babel', Jorge Luis Borges describes a menagerie of human beings who navigate the vast reaches of an infinite and eternal library: pilgrims, infidels, 'searchers', 'purifiers'. Many kiss the pages of its volumes in awed reverence; some scramble through its hexagonal galleries in a hunt for revelations and prophecies. Others, consumed by the promise of total knowledge, go mad.

What Borges did not imagine in his famous short story is a sect to rival all others, one that would respond to the rampant ubiquity of texts and accessibility of knowledge by using technology to turn inward and embrace a gratifying 'me now' solipsism: 'meformers'. Mor Naaman, Jeffrey Boase, and Chih-Hui Lai of Rutgers University introduced this neologism over a decade ago to characterise the majority of Twitter

users, who, in their analysis, privilege less information than 'meformation', posts about themselves and their feelings and beliefs.

The sound of Borges's 'feverish' library is the roar on the other side of a Google search box today. It can have us grasping for intellectual handles and footholds. At times it causes us to recoil from the exhausting complexity of the world to the comfort of our gut and intuition and to the acceptance of our tribe and circle of 'trust'. On social networks like Twitter and Facebook, we are invited to seek solace by generating 'meformation' – today the key commodity of what Shoshana Zuboff calls 'surveillance capitalism' – and by circulating posts and narratives and ideas that confirm our biases instead of exercising the muscles of our reason.

These digital technologies, which accelerate and proliferate information and harvest meformation, are enthusiastic handmaidens of disinformation. Our Disinformation and Media Literacy Special Interest Group (DML SIG) within the Trust and Technology SRI has sought to mobilise insights from social psychology and the humanities in the fight against disinformation, which we understand as a deliberate attempt to deceive and, in deceiving, to gratify.

Disinformation spreads because its frequent sensationalism can captivate, shock, amuse, and entertain. In a world whose boundless diversity is seemingly more evident than ever, disinformation can feed our sense of exception and envelop us in new feelings of belonging.

As our innovative partners at the Dutch team DROG frequently remind us, we cannot fact-check our way out of such a problem. Our DML SIG has accordingly aimed to fight fire with fire – to promote entertaining, even fun interventions against disinformation, from cultural activities and artistic events to free online games. The work of Jon Roozenbeek and Sander van der Linden has showcased the efficacy of such play in 'pre-bunking' disinformation and 'inoculating' players against its spread. We have also worked to advance translations of these interventions and tools across languages, from Arabic to Ukrainian, and to champion the idea of digital media literacy as a lifelong project and pursuit, as important and urgent to older

users as it is to children in school. Behind all of these efforts is a conviction that the humanities have a pivotal role to play in an interdisciplinary campaign to rethink our digital future(s). Culture mediates our relationships of trust and our relationships with technology. And much like Borges's library, it helps us envision alternatives of the present with warnings for the future.

Dr Rory Finnin is University Senior Lecturer (Associate Professor) in Ukrainian Studies at the University of Cambridge. He is the Founding Director of the Cambridge Ukrainian Studies programme. His primary research interest is the interplay of literature and national identity in Ukraine and the broader Black Sea region. Finnin is also co-convener of the University's Disinformation and Media Literacy Special Interest Group, a community of scholars and practitioners committed to advancing creative interventions against disinformation and 'fake news'.

Towards accountable algorithmic systems

Dr Jat Singh
Department of Computer Science and Technology

'Algorithm', once a term mostly used by those technical, has now become mainstream. The summer of 2020 saw students protesting against unfair algorithmic grade allocations, "f**k the algorithm" being one of their chants. That protest is but one example where people have pushed back against an algorithmic process. And as algorithms continue pervade and impact our lives, accountability will be increasingly on the agenda.

'Algorithmic systems' – colloquially describing systems with some technical or data-driven elements – are inherently *socio-technical*. Not only are people affected by these systems, but people are involved in their construction, operation and use. Views that technology is neutral, and that technology-related issues can be addressed (solely) through technical fixes are rapidly fading. Rather

now, we see 'algorithmic accountability' as a burgeoning area of interdisciplinary research, one that explicitly recognises that tackling issues in this context involves organisational, political, economic, and social considerations, not just those technical.

The *Compliant and Accountable Systems* Research Group[1] in the Department of Computer Science and Technology actively works in this space. The group focuses on the interplays of technology, law and policy, considering how technical and legal interventions might drive better compliance and accountability.

One key theme of our research is to bring about *meaningful* transparency, aiming at the information (and power) asymmetries between those leveraging algorithmic systems and those who oversee or are affected by them. Improving transparency won't, itself, solve these issues, but meaningful information about algorithmic systems can assist broader accountability regimes, by facilitating the understanding, oversight and scrutiny over systems and the parties involved.

Towards this, we have been developing the concept of *reviewability*,[2] which takes inspiration from the well-established principles of administrative law that govern public sector decision-making. Rather than focusing on the 'inner workings' of (or 'explaining') a system, reviewability seeks to enable a more holistic understanding of an algorithmic system. This is by providing a systematic framework for determining the relevant information – from the technical, organisational, and usage elements of a system – as is necessary for supporting meaningful oversight and scrutiny. In practice, reviewability entails recording details about an algorithmic system,

[1] www.compacctsys.net

[2] J. Cobbe, M.S.A Lee, J. Singh, *"Reviewable automated decision-making: A framework for accountable algorithmic systems",* ACM Fairness Accountability and Transparency (FAccT) 2021.

from before it's commissioned, right through its design, deployment, operation and use, as well as its resulting consequences.

Another key theme that we consider is what we call *rights engineering:* how algorithmic systems can better account for people's rights. Ensuring that systems accord with rights is an area of increasing attention; for example, work on issues of bias and fairness in AI often concern equality rights, while freedom of expression considerations are raised in the context of online content moderation.

In addition to how algorithmic systems impact rights, there are also considerations regarding how algorithmic systems support individuals in exercising their rights. For instance, the GDPR gives individuals certain rights regarding the processing of their personal data. However, this is an area under considered in practice – compliance in this space is patchy, the process of exercising one's rights can be cumbersome, and tensions exist between rights and other concerns such as security and privacy.[3] The research community is only just beginning to scratch the surface of the interplays between systems and rights; indeed, raising awareness seems an important step forward.

Algorithmic accountability, though a nascent area, is one rapidly growing in prominence. The above represents but a few examples of the many topics and challenges in this space. There is much to do to help make emerging algorithmic systems better work for us all.

[3] M. Veale, R. Binns, J. Ausloos, "*When data protection by design and data subject rights clash"*, International Data Privacy Law 2018; J. Singh, J. Cobbe, *"The security implications of data subject rights"*, IEEE Security & Privacy, 2019.

The Trust & Technology SRI 2018 - 2020

Executive Committee Trust & Technology SRI 2018-2020

Prof Simon Moore, Co-Chair
Department of Computer Science and Technology

Professor Simon Moore is a Professor of Computer Engineering in the Department of Computer Science and Technology at the University of Cambridge, where he undertakes research and teaching in the general area of computer design with particular interests in secure and rigorously-engineered computer architecture. Professor Moore is the senior member of the Computer Architecture research group.

Dr Jat Singh, Co-Chair
Department of Computer Science and Technology

Dr Jat Singh leads the Compliant and Accountable Systems research group in the Department of Computer Science and Technology. The group considers the intersections of computer science and law, exploring means for better aligning technology with legal concerns, and vice-versa. Jat is a Fellow of the Alan Turing Institute, the UK's national institute for data science and AI, and is active in the tech-policy space, having served on advisory councils for the UK Government and various regulators.

Dr Jennifer Cobbe
Department of Computer Science and Technology

Dr Jennifer Cobbe is a Research Associate in the Compliant and Accountable Systems Group, Department of Computer Science and Technology. Her research looks at the intersection of new and emerging technologies, law, and society from an interdisciplinary perspective. She is interested in legal responses, industry business models, and platform power; technical means for improving compliance and accountability of complex systems; and theoretical approaches to privacy, surveillance, and emerging technology. She is part of the Microsoft Cloud Computing Research Centre and a member of the Law Committee of the IEEE's

Global Initiative on Ethics of Autonomous and Intelligent Systems. Jennifer co-ordinated Trust & Technology activities in 2019/2020.

Dr Ella McPherson
Department of Sociology

Dr Ella McPherson is the Department of Sociology's Lecturer in the Sociology of New Media and Digital Technology as well as the Anthony L. Lyster Fellow in Sociology at Queens' College. She is also Co-Director of the Centre of Governance and Human Rights. Ella's research focuses on symbolic struggles surrounding the media in times of transition, whether democratic or digital. She currently focusses on human rights fact-finding in the digital age. Ella also leads The Whistle, an academic startup which aims to support the collection and verification of human rights information for evidence.

Dr Laura James
Department of Computer Science and Technology

As Entrepreneur in Residence at the Department of Computer Science and Technology, Laura helped to establish the Trust & Technology Initiative in 2018 and supported it part time during its first year, alongside other ventures working with emerging internet technologies in different contexts. She has worked extensively in technology and leadership roles in R&D, startups, civil society, humanitarian relief and co-operatives. Laura holds MA, MEng and PhD degrees in Engineering from Cambridge University, and is a Chartered Engineer.

Dr Andrea Lorenz
Department of Computer Science and Technology

Andrea has a hybrid academia-industry background, having swapped Medievalist scholarship for technology research in product development. An interest in translational research led to professional support for Life Sciences, Social Sciences and Computer Science at HE institutions. She served as co-ordinator of the Trust & Technology Initiative in 2020.

Steering Committee Trust & Technology SRI 2018-2020

Dr Anne Alexander
Centre for Research in the Arts, Social Sciences and Humanities

Dr Anne Alexander is Director of the Learning Programme at Cam-bridge Digital Humanities, a network of researchers at the University of Cambridge who are interested in how the use of digital tools is transforming scholarship in the humanities and social sciences. Her research interests include ethics of big data, activist media in the Middle East and the political economy of the Internet. She is a member of the Data Ethics Group at the Alan Turing Institute.

Dr Richard Clayton
Department of Computer Science and Technology

Dr Richard Clayton is a security researcher at the University of Cambridge and the Director of the Cambridge Cybercrime Centre, working in the field of work in the field of "security economics". He has research interests in email spam, fake bank "phishing" websites, and other Internet wickedness. As an expert in these areas, he is a regular speaker and media commentator. He has also assisted the APIG and APComms all-party groups of MPs in their inquiries into Internet issues, and he acted as "specialist adviser" for the House of Lords Science and Technology Committee's two inquiries into "Personal Internet Security".

Dr Robert Doubleday
Centre for Science and Policy

Dr Robert Doubleday has been Executive Director of the Centre for Science and Policy at the University of Cambridge since 2012. Previously Rob established CSaP's research programme. His re-search interests include the role of science, evidence and expertise in contemporary societies, in particular the relationship between scientific advice, public policy and democracy. Rob holds degrees in Chemistry (Imperial College, London) and Science and Technology Policy (SPRU, University of Sussex) as well as a PhD in Geography and Science & Technology Studies from University College London.

Dr David Erdos
Centre for Intellectual Property and Information Law

Dr David Erdos is Deputy Director of the Centre for Intellectual Property and Information Law (CIPIL) and University Senior Lecturer in Law and the Open Society in the Faculty of Law. He is also WYNG Fellow in Law at Trinity Hall. David's current research explores the nature of Data Protection in regards to the right to privacy, freedom of expression, freedom of information and freedom of research. This work intersects with debates on internet governance generally including, in particular, the liability and responsibility of "intermediary" actors such as Facebook and Google.

Dr Tanya Filer
Bennett Institute for Public Policy

Dr Tanya Filer leads the Digital State Project at the Bennett Institute for Public Policy. Her work focuses on GovTech (government technology) innovation ecosystems, and on digital government more broadly. Amid rapid technological change and deepening inequality, she seeks to understand how governments can better engage digital and emerging technologies, including for improved service provision and more meaningful forms of citizen participation. She also runs Tech States, the Institute's interview series featuring leading international voices on government and technology.

Prof Jennifer Gabrys
Department of Sociology

Prof Jennifer Gabrys is Chair in Media, Culture and Environment in the Department of Sociology at the University of Cambridge. She leads the Planetary Praxis research group, and is Principal Investigator on the ERC-funded project, Smart Forests: Transforming Environments into Social-Political Technologies. She also leads the Citizen Sense and AirKit projects, which have both received funding from the ERC. She writes on digital technologies, environments and social life. Recent publications are How to Do Things with Sensors (2019); and Program Earth: Environmental Sensing Technology and the Making of a Computational Planet (2016).

Dr Julian Huppert
Intellectual Forum, Jesus College

Dr Julian Huppert is the Founding Director of the Intellectual Forum, a new inter-disciplinary centre. His background is as a scientist, working on unusual structures of DNA using genomics to identify anti-cancer drug targets. He was the Member of Parliament for Cambridge, and Chair of the Panel of Independent Reviewers for DeepMind Health. He is now Deputy Chair of the NHS Cambridgeshire and Peter-borough CCG, a Director of the Joseph Rowntree Reform Trust, and a member of the Home Office Biometrics and Forensics Ethics Group, looking at big data and facial recognition.

Prof Adrian Kent
Department of Applied Mathematics and Theoretical Physics

Adrian Kent is Professor of Quantum Physics in the Department of Applied Mathematics and Theoretical Physics and a Distinguished Visiting Research Chair at Perimeter Institute for Theoretical Physics. His research interests span the foundations of physics and technological applications of quantum information. He has a strong interest in how we most effectively channel science and technological developments to shape our future in positive directions and to reduce catastrophic threats, and is a member of the scientific advisory board of the Cambridge Centre for the Study of Existential Risk.

Prof John Naughton
Centre for Research in the Arts, Social Sciences and Humanities

Prof John Naughton is a Senior Research Fellow at CRASSH, Emeritus Professor of the Public Understanding of Technology at the Open University, Director of the Wolfson Press Fellowship Programme and the Technology columnist of the London *Observer*. By background a systems engineer, he is an historian of the Internet whose research focusses on the network's impact on society. He was co-director of the *Technology and Democracy* and *Conspiracy and Democracy* research projects at CRASSH. His most recent work and publications have focussed on surveillance capitalism and the power and responsibilities of technology corporations.

Prof Daniel Ralph
Judge Business School

Daniel Ralph is Professor of Operations Research at the Cambridge Judge Business School, where he is also Academic Director of the Centre for Risk Studies. Professor Ralph is Fellow of Churchill College and a member of the Australian Mathematical Society, the Institute for Operations Research and the Management Sciences, and the Mathematical Optimization Society. He is Area Editor at Operations Research for Environment, Energy and Sustainability, and has served editorial roles in many other journals including Editor-in-Chief of Mathematical Programming (Series B).

Dr Manj Sandhu
Department of Medicine

Dr Manj Sandhu's research focuses on the integration of principles and procedures underlying population genetics and epidemiology. Together with current and emerging genome-wide technologies, this approach provides unparalleled opportunities to identify the biological mechanisms underlying the development of complex diseases and traits. He currently holds the post of Professor of Population Health & Data Sciences at Imperial College London.

Dr Simone Schnall
Department of Psychology

Dr Simone Schnall is the Director of the Cambridge Body, Mind and Behaviour Laboratory and Fellow of Jesus College. By combining insights and methods from social psychology and cognitive science her research explores how thoughts and feelings interact. She aims to understand how people make judgments and decisions about others, and about physical properties of the world. For example, her research has examined the role of bodily influences in the context of, first, moral judgments and behaviours, and second, perceptions of the spatial environment. In general the work seeks to understand why people often think and behave in seemingly surprising ways, and how to capitalize on insights from behavioural science to encourage adaptive choices in everyday life.

Dr Phillip Stanley-Marbell
Department of Engineering

Dr Phillip Stanley-Marbell is a University Lecturer in the Internet of Things. Phillip's research exploits the structure of signals in the physical world and the flexibility of human perception to make computation more efficient. He focuses on designing hardware architectures, algorithms, and programming language constructs that use an understanding of the physical world and the flexibility of sensing systems to improve the efficiency of computing systems that interact with nature. His research results range from fundamental theory, to algorithms, programming languages, and compiler tools. He frequently builds printed circuit board and FPGA prototypes to validate concepts.

Dr Adrian Weller
Department of Engineering

Dr Adrian Weller is Programme Director for AI at The Alan Turing Institute, and a Turing Fellow leading work on safe and ethical AI. He is a Principal Research Fellow in Machine Learning at Cambridge University, and at the Leverhulme Centre for the Future of Intelligence where he is Programme Director for Trust and Society. Adrian's interests span AI, its commercial applications and helping to ensure beneficial outcomes for society. He serves on several boards including the Centre for Data Ethics and Innovation. He is co-director of the European Laboratory for Learning and Intelligent Systems (ELLIS) programme on Human-centric ML, and a member of the UNESCO Ad Hoc Expert Group on the Ethics of AI.

Dr Jess Whittlestone
Centre for the Study of Existential Risk

Dr Jess Whittlestone is a Senior Research Associate at the Centre for the Study of Existential Risk. She works on various aspects of AI ethics and policy, with a particular focus on what we can do today to ensure AI is safe and beneficial in the long-term. She holds a PhD in Behavioural Science from the University of Warwick and a degree in Mathematics and Philosophy from Oxford University. She previously worked for the Behavioural Insights Team advising government departments on their use of data, evidence, and evaluation methods