AI reflections in 2020

We invited authors of selected Comments and Perspectives published in *Nature Machine Intelligence* in the latter half of 2019 and first half of 2020 to describe how their topic has developed, what their thoughts are about the challenges of 2020, and what they look forward to in 2021.

Anna Jobin

2 September 2019; Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019)

What was your Perspective about? Our article offered the first systematically conducted review of published artificial intelligence (AI) ethics guidelines. We analysed 84 documents and found that, despite an apparent convergence on certain ethical principles on the surface level, there are substantive divergences on how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented.

Do you feel the topic has developed over 2020? Scholarly and public discussions on AI ethics have certainly evolved. Although the illusion that 'ethical AI' is simply a technological matter still lingers, 2020 has seen an important push towards broader acceptance of the sociotechnicity of AI. Acknowledging the sociotechnical nature of AI systems requires us, as Pratyusha Kalluri put it succinctly[1], to centre less on fairness, or on 'AI for good', and more on power distribution and power differentials.

Has your own thinking on the topic evolved? Our overview of AI ethics guidelines has made clear to me that the devil is in the details. How meaningful is, for example, a pledge to 'human-centred AI' if there are no specifications as to how and by whom this will be defined, implemented, measured and controlled in practice? I have also realized that we should not let discussions about details make us lose sight of the big picture. For instance, it is crucial to pay attention to who gets to define the ethics of AI, and to the processes that decide what counts as ethical AI.

Were you surprised or worried by developments in AI in 2020? I was surprised to see students chanting 'f*ck the algorithm' in the streets of London, and I was excited to see protests against unjust algorithmic scoring having an impact. But I remain worried about how often AI is thrown at problems that cannot be solved by algorithmic systems. I remain worried about researchers and public-sector actors who are more concerned about their own status than their complicity with harmful structures and policies. I remain worried about the lack of whistleblower protection. I remain worried about big tech ignoring and suppressing critical voices and collective action.

Credit: AAUB/DigitalVisionVectors/Getty

What are your hopes or expectations for AI in 2021? Concerning the design, creation, training, deployment or use of AI, I expect the people and institutions that have decision-making power in these domains to prioritize the well-being of minorities and vulnerable communities. Overall, I hope to see a shift in how AI is governed, resulting in the allocation of more decision-making power to those who may or will actually be affected by these systems.

Kingson Man and Antonio Damasio

9 October 2019; Man, K. & Damasio, A. Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* **1**, 446–452 (2019)

What was your Perspective about? We proposed a new design principle for robots that would equip them with an analogue of feelings, which guides adaptive behaviour in living creatures. These 'vulnerable' robots are made of soft materials and are controlled by multi-sensory neural networks that can evaluate stimuli based on their consequences to homeostasis, the active maintenance of self-integrity.

Was there a specific motivation to write the article? Over the past few years we observed the remarkable advances occurring in AI on some perceptual and cognitive tasks. However, there is a concern in the community that these relatively narrow abilities will not generalize to other tasks, or even to the same tasks under real world complexity. Our ongoing research on the role of homeostasis and feelings in living creatures has shown that the response to feelings motivates creative and adaptive behaviour. We thought that it was time to import a similar mechanism or condition into artificial machines.

How has the topic developed over 2020? A major development in 2020 was OpenAI's GPT-3 language model, which demonstrates some truly astonishing text-generation abilities. But we believe that the discussion has returned to the same old debates on whether word co-occurrence statistics are sufficient to achieve understanding about the world. The field of embodied AI, which grounds knowledge in a vulnerable body's interaction with the world, remains under-appreciated. We predict that this will continue to be so, until an embodied AI has its own 'AlphaGo' moment and reaches, or exceeds, human abilities in a previously unthinkable domain.

What was the feedback to your article? We were surprised by the large number of reactions to our Perspective and, in particular, by the many responses that focused on the negative aspect of feelings; one would get the impression that our goal was to introduce 'fear and trembling' into robots. But we want to emphasize that the flip side of pain and suffering is pleasure and joy. And, we argued, the presence of any feeling at all unlocks abilities that are not possible in the absence of feeling. We are reminded of the anti-natalist argument that life is net suffering, and that, as a result, it is unethical to create new life. We reject this argument. We think that proliferation of machines of loving kindness could elevate humanity.

Has the COVID-19 pandemic affected your research? Yes and no. Everything is slower and yet more intense.

What are your hopes for AI in 2021? Investments in AI research are investments in future prosperity and security. We hope for a recommitment of government policymakers to the scientific principles of free exchange of ideas, open critique and debate, and respect for facts and expertise.

Georgios Kaissis and Rickmer Braren

8 June 2020; Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020)

What was your Perspective about? We outlined the emerging field of secure and private AI that employs a collection of innovative techniques to allow machine learning-based processing of sensitive or confidential data such as in medical imaging. These techniques can serve to train AI algorithms on larger datasets by combining multi-institutional data distributed all over the world and make these models accessible to more people without compromising patient privacy.

Was there a specific reason for you to write the article? We felt that privacy-preserving machine learning has reached a level of technical maturity that will soon permit a more widespread utilization for multi-institutional research. This is partly due to the availability of increasingly deployment-ready open-source software implementations such as OpenMined's PySyft or TensorFlow Federated for conducting research and creating products in this field. Furthermore, public opinion has been converging on higher awareness towards the value of protecting personal data for a long time. We hope that our article will motivate more researchers and the public to consider questions of privacy and invest into privacy-preserving methods for research and model development.

Do you feel the topic has developed over 2020? Definitely! The pandemic brought about a large-scale societal and political discussion about the ethics and the legal ramifications of automated contact tracing and data collection. In our view this highlights the importance of developing privacy-preserving tools in areas beyond medical imaging. Furthermore, privacy technology and research are evolving fast: new papers are published every day and 2020 saw new conferences, such as PriCon, featuring diverse and multi-faceted research from all over the world. The 2020 State of AI Report, widely considered a barometer of the industry, predicts a further rise in privacy-related research and deployment. Privacy preservation is now even being used as a marketing point in commercial products such as smartphones and wearables.

Did you get any surprising or useful feedback? We were overjoyed and humbled by the positive feedback from researchers in many different fields and by the great resonance from social media and other publications. We feel that this is mainly due to the fact that everyone can inherently identify with the requirement for privacy in such a sensitive field as healthcare and medicine. Many AI researchers also share our view that decentralized and privacy-preserving approaches will be the key to developing fair and representative algorithms on large and diverse datasets that, due to their private nature, can and should not be centrally shared.

Has your own thinking on the topic evolved? The field of privacy-preserving AI is full of positive innovation and constantly evolving. We are witnessing encouraging developments, for example dedicated hardware allowing cryptography and secure computation on handheld devices or new theoretical research into granular privacy tracking and budgeting in the field of differential privacy. The links between privacy-preserving deep learning and topics such as regularization or probabilistic inference are providing new insights on old questions and we feel that blockchain technologies are at a point of maturity where they can be employed alongside privacy-preserving systems for a variety of auxiliary tasks such as auditing.

Has the COVID-19 pandemic affected your research? We are privileged to have a robust infrastructure for online collaboration available to us at Technical University of Munich and Imperial College. OpenMined always has been a fully decentralized community and we were already successfully collaborating remotely before the pandemic. Therefore our productivity did not suffer much. We are saddened that fellow researchers in other fields and less privileged countries sustained large throwbacks in terms of productivity, research output and funding. We made efforts to combat social isolation and depression during the pandemic, and to offer an inclusive and welcoming climate for new team members and researchers. Our hearts and minds go out to all who suffer from or lost loved ones to this pandemic.

What are your hopes for AI in 2021? We hope that both academia and industry will continue on a value-aligned, innovation-driven course towards trustworthy AI development. We thus hope to see new breakthroughs beyond privacy-preserving AI, for example in verifiable AI, interpretability and — crucially — fairness, robustness, uncertainty quantification and reliability of AI-driven systems.

Julia Stoyanovich, Jay J. Van Bavel and Tessa V. West

13 April 2020; Stoyanovich, J., Van Bavel, J. J. & West, T. The imperative of interpretable machines. *Nat. Mach. Intell.* **2**, 197–199 (2020)

What was your Comment about? The main message of our article was the need to recognize the role of social psychology in building trustworthy algorithms. We argued that a research agenda on interpretability should answer three key questions: what are we explaining, for whom and for what reason?

Was there a specific motivation for you to write the article? Fairness and interpretability are top-of-mind for many data science researchers and practitioners, and that's a good thing. But we are sceptical that progress can be made without an understanding of how humans — including affected individuals, decision makers, data scientists and the public at large — perceive algorithms and their outputs. Our goal was to identify blind spots in the creation and communication of algorithms and to chart a path towards overcoming them.

Do you feel the topic has developed over 2020? When we started writing our article, there was already a growing awareness of these issues. Indeed, many people had pointed out cases of sexism or racism in various algorithms. Yet, there was no systematic understanding of what data scientists could do to increase trust in algorithms. We hope that our theoretical frameworks will generate more interdisciplinary collaboration on this issue, but have not seen much progress during this past year.

Did you get any surprising feedback? We received encouraging feedback, and are particularly happy to have heard from several students, who expressed an interest in working on the topics proposed in our article. We have also heard from practitioners, including human

resource executives who are tasked with removing bias in hiring and promotion procedures. We learned that there is often a misconception when it comes to promoting diversity in the workplace; namely, algorithms tend to be accepted as a welcome, unbiased alternative to human decision-makers who are regarded as biased. Our article has been eye-opening to some, and we are happy to see this conversation moving beyond academics to people who may not be aware of how algorithms are created and of all of the ways in which they, too, can be biased.

How has your own thinking on the topic evolved? We haven't changed our perspective, but we have certainly been surprised by how many people outside of academia are unaware of the issues we raised. Popular culture, at least in the United States, has created a conception of algorithms that is not fully grounded in reality, and so getting people to understand how algorithms are created is important.

Has the COVID-19 pandemic affected your research? Unfortunately, yes. It has forced us to prioritize our research on issues related to public health. Once the pandemic is over, we plan to refocus our efforts on the issues outlined in our paper. We are planning to obtain funding to formally test the ideas laid out in our paper. A silver lining is that, in the aftermath of the pandemic, the public is more keenly aware of both the potential benefits and the risks associated with large-scale data collection and analysis, of the importance of mitigating inequities in these systems, and of building trust. And some day in the near future, algorithms might be used to make important large-scale health decisions so this work might be particularly relevant.

What are your hopes for AI in 2021? We realize that AI is growing rapidly and represents a massive societal change. However, AI systems can be incredibly backward-looking, in large part because they are trained on historical data that by its nature represents the past. Therefore, we hope that programmers will think more deeply about the social and moral issues at play as they design future AI systems.

Brent Mittelstadt

4 November 2019; Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **1**, 501–507 (2019)

What was your Perspective about? Hundreds of public–private initiatives have been established across the globe since 2017 with the goal of defining common ethical principles and commitments to guide the future development and governance of AI. Common principles are a good start, but they are insufficient on their own to ensure we have ethical AI in the future. The difficult work of translating principles into practical guidelines, technical requirements, and inclusive policies will show us how far apart we actually are morally and politically in our notions of ethical AI.

Was there a specific motivation for you to write the article? I like to think the topic chose me rather than the other way around. The massive amount of time, effort and other resources that were being poured into AI ethics initiatives made the topic impossible to ignore. At the same time, I couldn't shake the feeling that we had collectively gotten ahead of ourselves in expecting universal consensus on what makes AI ethical, and how to achieve it in practice. Agreeing on common high-level principles is a good start, but as the core concepts, for example 'fairness' and 'transparency', can mean many different things in practice, consensus is hollow in practice. And so I was motivated by two closely related concerns. First, that these high-profile initiatives would start out strong, define a common high-level ethical framework, but then fail to support the substantial work needed to put principles into practice because of the time, effort and resources this work requires. And second, that disillusionment with AI ethics was just around the corner once it became apparent that principles can do very little on their own to fix AI's ethical problems. In short, I was worried that the hype around AI ethics had created an impossible expectation of ethics as a discipline.

Has your thinking on the topic evolved? I've seen the efforts that organizations and companies have put into translating principles into practice, for example by creating new ethics expert roles and procedures within organizations. These are positive steps, but I worry about the level of buy-in across all levels of organizations developing and using AI. In my current work I am looking at the feasibility of certification and licensing schemes to support both ethical and legal commitments in AI development and governance.

Has the COVID-19 pandemic affected your research? Not directly, but we've certainly seen ethics take centre stage in debates concerning COVID contact-tracing apps, immunity passports, and public health surveillance. In the development of digital contact-tracing apps in particular I've been disappointed to see how quickly public health interests have been dismissed in the name of privacy. It feels as though our (legitimate) concerns around privacy with the technology have been given absolute precedence over equally legitimate public health interests before we've had a chance to find the right balance between the two and translate this into the technology's design and governance.

What are your hopes for AI in 2021? My greatest hope is that we see sustained commitment to putting AI ethics into practice. In particular, I hope to see clear commitments made by organizations developing, procuring and using AI to specific forms of public transparency and third-party auditing. What is generally lacking in the field is a clear process to explain the difficult normative choices organizations are making around AI. Organizations should commit, for example, to public interfaces that allow people affected by their systems to request explanations of system behaviour. Similarly, they should explain how they define key concepts like fairness in practice, and how they arrived at these definitions. These processes need to be more accessible to regulators and to the people being impacted by AI.

Jason Eshraghian

9 March 2020; Eshraghian, J. K. Human ownership of artificial creativity. *Nat. Mach. Intell.* **2**, 157–160 (2020)

was your Perspective about? Advances in generative algorithms have enabled neural networks to generate synthetic datasets, from photorealistic videos to human-like text. But when the creative process is automated by a programmer, in a style determined by the trainer, using features from public and private datasets, who is the proprietary owner of the rights in AI-generated artworks and designs? My Perspective seeks to answer this question from a legal standpoint and establishes four guiding principles that users of creative-AI can apply to ensure their own interests are protected.

Was there a specific motivation for you to write the article? There was much debate about who owns the rights to AI-generated artworks following the high-profile sale of the portrait *Edmond de Belamy* (for US$432,500), which had multiple contributors to its development. My article was written with the intention to address the uncertainty in legal principles surrounding generative AI. By providing a set of principled guidelines, users of generative

AI can be more confident in displaying, distributing or commercializing their work without the risk of infringing upon the rights of others.

Do you feel the topic has developed over 2020? It remains an open issue. The case of Author's Guild versus Google set a precedent that allowed Google to train its database on copyrighted books to develop their Book Search algorithm. This indicates that training a discriminative model on copyrighted material is perfectly legal and has the potential to be applied to future legal challenges on generative models. But it remains to be seen how protection can be granted to AI-artwork itself, and whether this will vary between jurisdictions.

Has your own thinking on the topic evolved? AI is thought to diminish human involvement in the creative process. But I have found that, in a way, it fosters collaboration between many different parties and that there is still a substantial human element behind artificial creativity. Multiple people are contributing to public repositories, and the artworks of thousands of people are pooled together in a training set. While technology may be lessening the scope for human involvement in the creative process, it comes with the possibility for thousands of creators to contribute towards a single piece of artwork.

Were you worried by other developments in AI in 2020? 2020 has shown that dataset bias is a problem that goes far beyond the data. The inconsistent access to technology, healthcare and education across nations, races and socioeconomic standing will mean that most training data presently available are not an accurate representation of a population. As a result, these networks will be entrenched with bias. They will be skewed to favour those who have had historical access to these basic essentials. Active steps must be taken to eliminate such bias, and the limited recognition of the problem going beyond the data is concerning.

Has the COVID-19 pandemic affected your research? I am all of a sudden a hero for continuing to live my life indoors, spending 16 hours a day on a computer.

What are your hopes for AI in 2021? I hope 2021 will bring greater focus on the pressing issues surrounding AI in society, such as dataset bias, data privacy and the carbon footprint associated with the extremely large networks being trained, and a shift away from the distant and nebulous concepts that saturate many discussions, such as technological singularity.

Marta R. Costa-jussà

14 October 2019; Costa-jussà, M. R. An analysis of gender bias studies in natural language processing. *Nat. Mach. Intell.* **1**, 495–496 (2019)

What was your Comment about? Gender bias studies in natural language processing (NLP) have shown progress in bias detection, evaluation of AI systems in terms of bias and even in algorithmic bias mitigation. However, to make an impact on society, the field needs a clear path forward. My Comment asked whether current studies offer sufficient conceptualization of the bias challenges and whether there exists a clear joined-up effort.

Was there a specific reason or motivation for you to write the article? At the time of writing the article, I was pleased that the research field had started to pay attention to the bias problem. However, each publication was working in a separate direction and efforts were not directed sufficiently to suggesting solutions for the problem. As a consequence, the field was producing a large amount of quantitative work with complex maths, but lacking consistency and was far from having a social impact.

Has the topic developed over 2020? A related paper was published this year that analysed in detail 146 papers dedicated to gender bias in NLP[2]. The authors concluded that motivations are often vague and that there is a lack of conceptualization of bias. This led to the initiative of a 'bias statement' in the Workshop of Gender Bias in NLP: authors were encouraged to give explicit consideration to the wider aspects of bias. This resembles other initiatives, for example, the 'impact statement' at NeurIPS where authors are encouraged to discuss potential broader consequences of the work, in terms of ethical and future societal aspects.

Do you have new insights on the topic? At first I mainly assumed that the bias present in our algorithms comes from the data. Now, I think that we are able to mitigate this bias by improving our algorithms, for example, in terms of generalization[3]. My experience is that changing the data with automatic procedures is quite difficult, and it is even more challenging to do this non-automatically, since this implies a societal change. Therefore, I think that work on algorithms that are able to detach from the data (that is, generalize better) may help in what should be our main goal: producing system outputs without biases.

Has the COVID-19 pandemic affected your research? The productivity of the research group I am co-leading seems to have increased and so have the interactions with international teams, as we've been forced to work in virtual mode. At the same time, I feel that not interacting face-to-face is diminishing creativity and potential new co-operations within the group.

What are your hopes for AI in 2021? The AI community is little by little becoming more inclusive and fairer thanks to several initiatives such as specific workshops and new policies at conferences. For 2021, it will be absolutely necessary that social and educational policies from governments align with the long-term challenge of combating bias. Moreover, I expect that a higher proportion of scientific contributions and social policies will explicitly take into account these topics by deeply questioning our data and algorithms, rethinking hierarchies, challenging power or embracing pluralism[4].

Asaf Tzachor

22 June 2020; Tzachor, A., Whittlestone, J., Sundaram, L. & Ó hÉigeartaigh, S. Artificial intelligence in a crisis needs ethics with urgency. *Nat. Mach. Intell.* **2**, 365–366 (2020)

What was your Comment about? We argued that emergencies, such as the pandemic, engender circumstances in which AI technologies may be deployed at extraordinary speed, scale and depth, but at the expense of adequate oversight, foresight or rigorous risk assessment. We suggested several pre-emptive approaches to ensure safe, secure and ethically sound use of AI in crises. For instance, we recommended designing for transparency and explainability, and using adversarial techniques such as 'red teaming'.

Was there a specific motivation to write the article? Our immediate motivation was that we were concerned about the potential misuse of private data in attempts to monitor the spread of the disease. We were also concerned about the possibility that some governments may depart from their liberal tradition and exploit the abnormal circumstances to keep personal data in perpetuity, and to legitimize algorithm-assisted instruments of coercion.

A more profound motivation was given by the fact that, presumably, the pandemic is the latest but not the last in a line of emerging zoonoses[5]. In preparing for future

public health crises, we sought to call the attention of data scientists, data engineers, ethicists and healthcare specialists to the difficulty of maintaining ethics in a crisis.

Did you get any surprising feedback? We were very pleased that our article went on to inform a number of task forces and parliamentary committees on AI and its implications for the coronavirus crisis.

Has your own thinking evolved? My thinking on the topic of 'AI ethics in crisis conditions' has extended beyond the pandemic; the insights we gain from the analysis of AI performance in the current emergency are pertinent and transferable to predictable adversities in other domains. Namely, I see great urgency to devise participatory and value-sensitive AI design roadmaps in preparation to cope with extreme weather anomalies resulting from anthropogenic climate change. This is an area of predictable disasters where algorithms, in combination with other machines, can play a pivotal role in prediction, preparedness, rapid response and optimal allocation of scarce life-supporting resources.

What AI developments in 2020 were you excited by? I was excited to see efforts to apply machine learning techniques to disaster risk reduction. One inspiring endeavour is the Google Flood Forecasting initiative[6], which partners with the Bangladesh Water Development Board and the International Red Cross, to improve the spatial and temporal accuracy of flood forecasting and real-time notifications. Another promising development is the Descartes Labs Platform employing machine learning methods to enhance prediction accuracy of wildfires occurrence and progression.

Has the COVID-19 pandemic affected your research? The pandemic affected our research in several ways: we were obliged to suspend community-based fieldwork in a selected set of developing countries assessing socio-cultural barriers to AI adoption in safety-critical domains. At the same time, the pandemic presented us with a stress test of our institutions and best available technologies.

What are your hopes for AI in 2021? I hope 2021 will see advancements in data stewardship across safety-critical domains, spanning more countries. Mainly, data portals should give visibility to marginalized communities, and data managers should ensure their datasets are discoverable, accessible and reusable, and can be easily aggregated and interpreted. A noteworthy initiative in this regard is the Whose Knowledge? campaign to decolonize the Internet's languages. In the same vein, I hope vulnerable populations — whether elders at risk of respiratory illness or disadvantaged rural communities at risk of natural disasters — gain greater access to intelligent decision support systems in times of a crisis. Multidisciplinary scholarship will be vital to attain these goals, and so 2021 should hopefully see scholars from diverse disciplines, social groups and cultures engaging with these issues.

Aimun A. B. Jamjoom

13 April 2020; Jamjoom, A. A. B., Jamjoom, A. M. A. & Marcus, H. J. Exploring public opinion about liability and responsibility in surgical robotics. *Nat. Mach. Intell.* **2**, 194–196 (2020)

What was your Comment about? As robotic systems become more autonomous, it gets less straightforward to determine liability when humans are harmed. In our article, we discussed this emerging challenge in the context of surgical robotics and introduced the iRobotSurgeon Survey, which aims to explore public opinion towards the issue of liability with robotic surgical systems.

Was there a specific motivation for you to write the article? In the past few years, machine learning advances have enabled the development of increasingly autonomous robotic systems. These advances show that a future in which a patient undergoes surgery by a robotic surgical system with minimal supervision from a human surgeon is no longer a matter of science fiction. However, this shift in decision-making from humans to autonomous systems poses a legal challenge in determining liability. We believe that there is a need to explore public attitudes to these questions and developed the iRobotSurgeon Survey. The survey presents five hypothetical scenarios where the patient comes to harm and the respondent needs to determine who they feel is mostly responsible: the surgeon, the robot manufacturer, the hospital or another party. The motivation behind our Comment was to provide the rationale and background for the survey. In particular, we wanted the article to help raise awareness about the issue and encourage engagement with the survey.

Has the COVID-19 pandemic affected your research? The iRobotSurgeon Survey was launched at the start of 2020 just as the COVID-19 pandemic started to take off. We had been planning to promote the survey through both the mainstream and social media, but had to postpone as attention focused on the pandemic. After this initial delay, we have been able to promote the survey and have gathered over 1,400 responses from 60 countries around the world.

Do you have any specific hopes for AI in 2021? In 2021, we hope that the question of liability in autonomous systems becomes a growing area of interest in the AI community and regulators. In particular, we would like to see more research on societal expectations and desires on how these systems should be regulated and on what legal frameworks could underpin these developments. Importantly, interdisciplinary collaboration between technologists, ethicists, lawyers, surgeons and patients will be vital to building consensus on how liability is ascribed as decision-making shifts from surgeons to surgical robotic systems.

Mariarosaria Taddeo

11 November 2019; Taddeo, M., McCutcheon, T. & Floridi, L. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nat. Mach. Intell.* **1**, 557–560 (2019)

What was your Perspective about? We argued that trustworthy AI is a misnomer because the inherent lack of robustness of AI makes it impossible to assess its trustworthiness. We focused on the use of AI for cybersecurity purposes and considered the risk that trusting AI in this domain would pose. We suggested that governance of AI for cybersecurity purposes should aim at deploying reliable rather than trustworthy AI.

Was there a specific reason for you to write the article? At the time of writing the Perspective, several governments around the wold explicitly mentioned the use of AI to improve the security of critical national infrastructures, such as transport, hospitals, energy and water supply. In the two years before that, a number of national and international initiatives to foster ethical governance of AI were published, sharing a central focus on the concept of trust. They assumed uncritically that trust in AI is a necessary element to foster its uptake. We considered this misleading when focusing on AI in general and dangerous when considering AI applications in cybersecurity in particular. In the Perspective, we used the definition I previously introduced for trust — a second-order property that is qualified by the delegation of a task and the lack of monitoring over the way in which

the task is performed[7] — to distinguish trust from reliance, which envisages some form of control over the execution of a given task.

Has your own thinking on the topic evolved? The opportunities and challenges linked to the adoption of AI in cybersecurity made me consider the relation between trust and innovation, in particular digital innovation. Once adopted, digital technologies become an interface through which we interact, change, perceive and understand others and our environment. These technologies blend in the 'infosphere'[8] to the point of becoming an invisible interface[9], one that we are encouraged to trust and which we may easily forget about, at least until something goes (badly) wrong. This 'trust and forget' dynamic is problematic, because it may lead to the erosion of human control on the impact that digital technologies have on our societies. However, the picture is not all bleak. It becomes clear that citizens' trust is not placed in the technology but in the public institutions deciding on, and governing, its deployment. Citizens trust institutions to oversee the deployment of AI systems that are safe, reliable, have appropriate levels of transparency, are monitored appropriately, and are accompanied by accountability procedures and redressing measures for any unwanted consequence following the use of AI.

What are your hopes for AI in 2021? I am optimistic that key lessons in the area of digital ethics can be learned from the pandemic. Our lab views research on the conceptual, ethical, legal and social implications of digital technologies as essential groundwork to inform policies for the governance of these technologies. As in many corners of the world public agendas identify digital technologies as a key element to design post-pandemic societies, I hope to contribute to informing these initiatives, with the goal to leverage digital technologies to design democratic, pluralist, sustainable societies.

Edoardo Sinibaldi

11 May 2020; Sinibaldi, E. et al. Contributions from the Catholic Church to ethical reflections in the digital era. *Nat. Mach. Intell.* **2**, 242–244 (2020)

What was your Comment about? Digital innovation and technological progress increasingly affect our vision of humanity, as key concepts such as embodiment, agency and intelligence are stretched to apply to machines. The Catholic Church feels the responsibility, as part of its mission, to nurture global cooperation and inclusive dialogue on machine ethics, through scientific events and journal publications.

Was there a specific motivation for you to write the article? With the growing impact of artificial intelligence, all parts of society must be mobilized. We were motivated to contribute to the interdisciplinary discussions. In particular, we felt that the time was ripe to interweave faith with science and technology, with the aim to identify a path where we respectfully listen to different voices on the topic of machine ethics.

Do you feel the topic has developed over 2020? The pandemic has sped up discussions about the impact of AI in society, due to a boost in diffusion and implementation of digital technologies, and an increasing exchange of data. We now have a clearer understanding of opportunities and risks; for example, increased knowledge and data sharing can be beneficial in a pandemic but carries a risk of unwarranted surveillance and control. Despite the urgency of the current situation, we should consider implications for both short and long term.

Did you get any surprising or useful feedback? We have received many positive comments about the value of bridging scientific and technological research with the Church reflection on machine ethics. At the institutional level, the Rome Call for AI Ethics has elicited substantial interest. In September, the Food and Agriculture Organization of the United Nations (FAO) amplified the Call through an international conference, which featured concrete AI applications for the promotion of sustainable ways to achieve food and nutrition security. Furthermore, in October, the Sapienza University of Rome, which is one of the largest in Europe and oldest worldwide, has been the first academic institution to sign the Call (further universities are expected to sign in the coming months).

What development in AI in 2020 were you excited by? As we highlight in the Comment, religious denominations at large, as full participants in pluralist societies, should be part of an inclusive dialogue. Therefore, the Pontifical Academy for Life (PAV) started a networking project, and we are now excited by the growing possibility to globally collaborate with other denominations in order to increase societal awareness and responsibility on AI and machine ethics.

Has the COVID-19 pandemic affected your research? Yes, indeed. On the one hand, our research efforts were steered towards key applications, such as (ethical) processing of healthcare data and AI- and robot-assisted remote support in hospitals and clinics. On the other hand, the pandemic has revealed that we were only partially ready. This highlights a need to carefully consider the global impact of our research efforts in a connected society, on short and long terms. In a broader perspective, the pandemic is urging us to intertwine research and solidarity more strongly.

What are your hopes for AI in 2021? We hope that efforts fostering ethical reflections on machine intelligence will unite many parts of society and that shared principles are turned into actions. We hope that everyone involved in the development, deployment and use of AI technologies will take responsibility to pursue respectful and equitable applications, by firmly keeping humans at the centre.

Yipeng Hu

22 May 2020; Hu, Y. et al. The challenges of deploying artificial intelligence models in a rapidly evolving pandemic. *Nat. Mach. Intell.* **2**, 298–300 (2020)

What was your Comment about? We highlighted two of the main challenges for successful adoption of AI models in the fight against the pandemic. The first challenge is that clinical needs are moving as the epidemic progresses. For example, what could be a useful AI model at the peak of the pandemic might be very different than what is required at the beginning. The second challenge is the necessity to translate models to local healthcare situations, for which we suggest a local adaptation strategy.

Do you feel the topic has developed over 2020? By and large, the challenges remain, as is evident from the fact that very few AI models for diagnosis and prognosis are successfully translated into clinical practice and capable of helping patients today.

Has your own thinking on the topic evolved? Only a few months into the pandemic, many papers appeared on initiatives in applying AI to help in some way. However, I increasingly recognized that a substantial effort is needed to ensure that such initial developments are of sufficient quality to be of use in further research developments. A high level of scientific rigour is required, such as regarding the training and validation data, patient cohort representativeness, experiment design and statistical analysis, to realize the potential

clinical value of the reported results. Several prospective studies have already started, such as in AI applied to medical images for diagnosis and prognostications, but these are developed at an unprecedented speed compared to any previous AI algorithm development.

Has the COVID-19 pandemic affected your research? We are now in a difficult position to get clinical data. In particular, surgical and interventional applications usually require engineering researchers to acquire data during those procedures, but this is not possible now. Lab experiments that require more than one person are also being discouraged.

Do you have any specific hopes for AI in 2021? Yes — more clinical translation of AI models, not only to help in the pandemic, but for many other clinical areas!

Miguel Luengo-Oroz

22 May 2020; Luengo-Oroz, M. et al. Artificial intelligence cooperation to support the global response to COVID-19. *Nat. Mach. Intell.* **2**, 295–297 (2020)

What was your Comment about? There are hundreds of multidisciplinary AI research initiatives at molecular, clinical and societal scales that can help fight COVID-19. For AI to make a real impact and to overcome a hyper fragmented space with limited operational deployments, we need digital cooperation and solidarity across borders and stakeholders, including responsible and scalable approaches for data, models and code sharing. We also need mechanisms for adaptation of applications to local contexts and priorities.

Was there a specific motivation for you to write the article? The COVID-19 crisis is not just a public health crisis but affects every other socio-economic dimension as it disproportionally affects vulnerable populations and potentially exacerbates inequalities. Addressing this challenge requires collaboration between disciplines and communities. Moreover, many AI researchers did not know where to start and how efforts could be most effective. In our Comment, we wanted to provide a framework to think about the big picture of how AI can help against the pandemic. We also wanted to motivate a greater cooperation between domain experts (policymakers and healthcare professionals) and the AI community to responsibly build effective and scalable solutions.

Has the COVID-19 pandemic affected your research? Since February, our team has shifted priorities and has been working together with other United Nations agencies including the World Health Organization and the UN High Commissioner for Refugees (UNHCR) to support the COVID-19 response. We have landscaped AI and COVID-19 applications, worked on infodemics research roadmaps, and mapped inequalities and privacy risks that may arise from the use of AI applications in the pandemic. Furthermore, our team has been working with public and private sector partners in multiple data-driven operational projects such as supporting local teams to counter health misinformation in the Global South or creating epidemiological models to understand the potential impact of public health interventions in refugee camps and settlements.

What are your hopes for AI in 2021? My hope is that science — including AI developments — and solidarity will inform policy-making more directly during the pandemic response. For instance, I hope to see creative and local public health interventions guided by the next generation of precision epidemiology based on massive computational simulations with big data in time and space for detailed predictive modelling. Next year will be critical in the fight against the infodemic. Trolls and conspiracy theorists will continue to attempt to undermine the epidemic response and the vaccination roll-out. I expect social media companies to finally be required to change some of the core assumptions, including how to optimize and trade for people's attention. From the AI perspective, this pressure might stimulate new ideas around recommendation systems that do not lead to echo chambers and rabbit holes, and around collaborative systems against the proliferation of hate and anti-vaccine speech. Besides the pandemic, the climate crisis should also be a wake-up call for the AI community in 2021. I hope to see proposals for energy labelling in AI models as potential industry standards. ❐

Anna Jobin[1], Kingson Man[2], Antonio Damasio[2], Georgios Kaissis[3,4,5], Rickmer Braren[3], Julia Stoyanovich[6,7], Jay J. Van Bavel[8,9], Tessa V. West[8], Brent Mittelstadt[10,11], Jason Eshraghian[12], Marta R. Costa-jussà[13], Asaf Tzachor[14], Aimun A. B. Jamjoom[15], Mariarosaria Taddeo[10,11], Edoardo Sinibaldi[16,17], Yipeng Hu[18,19,20] and Miguel Luengo-Oroz[21]

[1]STS Lab, University of Lausanne, Lausanne, Switzerland. [2]Brain and Creativity Institute, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA. [3]Department of Diagnostic and Interventional Radiology, Faculty of Medicine, Technical University of Munich, Munich, Germany. [4]Department of Computing, Imperial College London, London, UK. [5]OpenMined. [6]Department of Computer Science and Engineering, Tandon School of Engineering, New York University, New York, NY, USA. [7]Center for Data Science, New York University, New York, NY, USA. [8]Department of Psychology, College of Arts and Sciences, New York University, New York, NY, USA. [9]Center for Neural Science, New York University, New York, NY, USA. [10]Oxford Internet Institute, University of Oxford, Oxford, UK. [11]The Alan Turing Institute, British Library, London, UK. [12]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. [13]Universitat Politècnica de Catalunya, Barcelona, Spain. [14]Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK. [15]Centre for Clinical Brain Sciences, Edinburgh University, Edinburgh, UK. [16]Italian Institute of Technology, Genoa, Italy. [17]Working Group on Roboethics of the Pontifical Academy for Life, Vatican City, Vatican City. [18]UCL Centre for Medical Image Computing, University College London, London, UK. [19]Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK. [20]Department of Medical Physics and Biomedical Engineering, University College London, London, UK. [21]United Nations Global Pulse, New York, NY, USA.

Published online: 19 January 2021
https://doi.org/10.1038/s42256-020-00281-z

References

1. Kalluri, P. *Nature* **583**, 169 (2020).
2. Blodgett, S. L., Barocas, S., Daumé, H. III & Wallach, H. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 5454–5476 (ACL, 2020).
3. Webster, K. *Google AI Blog* https://ai.googleblog.com/2020/10/measuring-gendered-correlations-in-pre.html (2020).
4. D'Ignazio, C. & Klein, L. F. *Data Feminism* (MIT Press, 2020).
5. *Preventing the Next Pandemic: Zoonotic Diseases and How to Break the Chain of Transmission* (United Nations Environment Programme, 2020).
6. Nevo, S. *Google AI Blog* https://ai.googleblog.com/2019/09/an-inside-look-at-flood-forecasting.html (2019).
7. Taddeo, M. *Minds Mach.* **20**, 243–257 (2010).
8. Floridi, L. *Ethics Inform. Technol.* **4**, 287–304 (2002).
9. Taddeo, M. & Floridi, L. *Science* **361**, 751–752 (2018).