From the Ground Truth Up: Doing AI Ethics from Practice to Principles

Abstract

Recent AI ethics has focused on applying abstract principles to practice. This paper goes the other way. Starting from the experiences of AI-intensive companies, knowledge is produced and transferred upward to influence theoretical debates surrounding these questions: 1) Should AI as trustworthy be sought through explainability, or accurate performance? 2) Should AI be considered trustworthy at all, or is reliability a preferable aim? 3) Should AI ethics be oriented toward establishing protections for users, or toward catalyzing innovation? Specific answers are less significant than the larger demonstration that AI ethics is currently unbalanced toward theoretical principles, and will benefit from increased exposure to real world practices.

Keywords

AI ethics, Trustworthy AI, Philosophy and AI, Medical AI, Case studies

Words

4200 (Text only)

1. Introduction

Artificial intelligence is defined as knowledge produced from pattern recognition, which contrasts with the Kantian vision of human knowledge as created by sequential reasoning. This distinction at – and as – the source of understanding means AI cannot be relied upon to obey human conventions of rationality. So, an ethics is needed to domesticate the machines.

Initial work has been largely theoretical, with 84 sets of principles introduced in the last several years, and more on the way. They accumulate along with a challenge. Morley, Floridi, Kinsey and Elhalal (2020) write that these abstract principles urgently

require effective translation for application in lived reality. While their work contributes to the translating, it also implies complementary explorations circulating in the other direction: instead of progressing from principles to practice, they *start* from tangible human experience and only subsequently transfer up to abstract theory. This commentary pursues them. It joins contemporary AI ethics by seeking to unite principles and practice, but it diverges by working from the ground up.

2. Cases

Going from practice to principles starts with cases. The two discussed here were produced by a team of philosophers, computer scientists, lawyers, and doctors organized [Redacted]. Working with real startup companies, we collaboratively explore their development experience and then react with a report from the group's diverse members, with attention split between ethics, technology, law, and medicine [Redacted].

Our skin lesion case started with a team led by Andreas Dengel and their solution to a debility currently afflicting artificial intelligence diagnoses of skin cancer. Typically, a skin lesion image is analyzed by a neural network to predict whether the lesion is malignant. The procedure is noninvasive, efficient and, in terms of accuracy, machines are now outperforming well-trained dermatologists (Brinker et al. 2019), but the technology's use nevertheless remains limited. The obstacle is the AI blackbox (Lucieri et al. 2020). Doctors may be convinced that the image analysis *generally* works better than their own eyes and experience, but because there is no way to know how the machine reached its conclusion, they fear their patient may be an outlier. The hesitation is understandable: advances in image recognition technology have been accompanied by startling errors. There is even a narrow research area dedicated to provoking comedically wrong outputs, like bananas mistaken for toasters (Brown et al. 2017). The barrier, that means, to AI-fortified skin health is not technological advance so much as doctors' confidence, and that apprehension converted into a business opening for Dengel and his team. Their Explainable AI in Dermatology product – exAID – wraps around existing artificial intelligence diagnoses and translates the AI method into traditional medical language and reasoning. Once the AI processing is explained, dermatologists may confidently confirm or reject the mechanical diagnosis. Technology that was accurate but neglected now becomes practically useful.

The other case starts from cardiac arrest in Denmark, and with a team lead by Stig Nikolaj Blomberg (Blomberg et al. 2021). They responded to an urgent question: Could AI eavesdrop on frantic 112 calls – the Danish 911 – and perceive humanly imperceptible clues that the subject was suffering cardiac arrest as opposed to some

less urgent malady? The information is crucial because cardiac arrest requires specific and immediate treatment, even from bystanders, for survival as every minute without resuscitation increases fatality probability by about 10% (Murphy et al. 1994). Dispatchers at Denmark's Emergency Medical Center were failing to identify 25% of the incoming cardiac arrest calls, and so losing the precious opportunity to provide the caller with instructions in cardiopulmonary resuscitation (Blomberg et al. 2019). To save lives, Blomberg's group developed a machine learning tool that could detect cardiac arrest, and alert Dispatchers with a light added to their console. When the technology was implemented, it produced unsurprising and also surprising results. Unsurprisingly, true cardiac arrest was recognized more frequently and quickly by the artificial intelligence than by its human partner [Redacted]. However, the machine was also less specific: the AI-returned many false positive alerts which were largely ignored by the dispatchers. This leads to the surprising result. The dispatchers also largely ignored the true positive alerts. They ignored the AI almost entirely. Dejectedly, Blomberg concluded: "While a machine learning model recognized a significantly greater number of out-of-hospital cardiac arrests than dispatchers alone, this did not translate into improved cardiac arrest recognition by dispatchers" (Blomberg et al. 2021)."

3. Explainability or Performance?

At the highest level, European AI ethics is dedicated to creating technology that is trustworthy (HLEG 2019). Both cases confronted that trust problem, but the project leaders diverged on a question underneath: *why* believe in algorithmic conclusions in the first place? This uncertainty is not so much about how much confidence exists, but on what the confidence is built. The distinction divided the teams fundamentally.

Dengel's exAID product builds trust from explainability: machine image analysis will be accepted when it is *understood*. Consequently, his AI wrapper is built to show how skin images are classified, and to verify that the processing functions through disease-related concepts similar to those employed by dermatologists. The key is to translate away from the statistics and probabilities, and into seven clinical skin characteristics perceived through close visual inspection. They are: Typical Pigment Network, Atypical Pigment Network, Streaks, Regular Dots and Globules, Irregular Dots and Globules, Blue Whitish Veil (Lucieri et al. 2020, Argenziano et al. 1998). The presence of these telling traits is quantified and then overlayed on the image of the skin lesion under scrutiny. Significantly, it is not that the numbers are outputted to *replace* the images with an objective result, instead, they *describe* the images, they direct doctors' attention back to the visual evidence so they can check for themselves. Without the exAID wrapper, all a dermatologist receives is a cold numerical probability sent back

in return for submitted images. With the wrapper, the statistics no longer substitute skin pictures, they help doctors see the pictures more clearly.

In essence, machine learning reverses: instead of the material world converting into a digital score, the score guides a way back into material experience. Numbers serve eyes, not the other way around.

The original skin lesion picture also receives a second overlay, a layer of color indicating exactly where the AI detected the information resulting in diagnosis. So, the doctor learns not only of a blue whitish veil, but where on the lesion it can be found. Here again, explaining the AI does not mean adding still more digital information or another set of statistical methods (Shapley values or similar) to approximate which pieces of data contributed how much to prior statistical processing (Chen 2021). Instead, it means translating the experience of interacting with the system into the familiar and human modes of seeing, locating, touching. The machine is humanized, anthropomorphized.

In Denmark, a different strategy: instead of understanding, Blomberg and his team leveraged power. AI performance *imposed* trust. To force dispatcher respect for the cardiac arrest alerts, the machine was tuned to defeat humans. While it is true that an abundance of false positive alerts got logged along the way, the hard fact remained that true cardiac arrest was detected faster by algorithms and data than by human listening and experience.

The idea of helping the human dispatchers catch up with the AI instead of leaving them behind – perhaps by developing software to clarify or augment the audio the dispatchers heard – was never broached in discussions with our group. Just the opposite, we learned that as the AI development progressed, humanity reduced. Originally, the AI processed the raw audio of calls, thick with their anguish. However, the discovery was made that background screams and lamentations were major sources of false positive alerts, and so a two-stage approach was developed. An initial filter eliminated human emotion by transcribing the calls into dry words, and then a second process analyzed the language for patterns in vocabulary, in sentences, in questions and answers, and in specific, described characteristics. Blomberg explained that if the caller states that the subject is unconscious, then the probability of cardiac arrest rises. If blue lips are mentioned, the probability also rises. If both, the alarm illuminates [Redacted]. However, beyond that and a few similar anecdotes, there was no human-oriented discussion of the AI process, nothing that would make sense to a doctor. There were only the statistical outcomes of sensitivity (the detection of cardiac arrest) and

specificity (the ratio of true cardiac arrests detected, against false alerts), and how they could be improved.

Though the model was partially open-sourced (Havtorn et al. 2020; Maaløe et al. 2019), our group did not pursue a technical understanding of explaining how the system worked because we were convinced that winning human trust for the AI decisions would originate and remain within the parameters of performance defined as speed and accuracy. Ultimately, the objectivity – the mathematical certainties – were not just descriptions of functionality but conceptual rigidities that commanded human respect by humiliating subjective and uncertain natures. While emergency call dispatchers considered and doubted and floundered and let seconds drip through their indecision, hard numbers responded. The result for the emergency call AI technology was a kind of trust not won from dispatchers so much as stamped onto them. As opposed to the previous case which drew human explanations from an inanimate machine, here, the power of inanimate machines was programmed to crush human doubts.

In actual practice in Copenhagen the dispatchers were not crushed, they resisted by ignoring the technological prompts. But that failure does not change the nature of the strategy, it only requires still more engineering and perfecting.

Which of the two paths to trustworthy AI is recommendable? Is it explainability so the machine will be *understood* well, or accurate performance so the machine will *work* well? The case for explainability flourishes in German doctors' offices: it was *because* the skin analyzing machine was finally understood well, that it was allowed to work well.

In the Danish emergency call center, a different choice was unavoidable. Faced with the reality that dispatchers are ignoring cardiac arrest alerts, Blomberg tuned his machine still tighter, and pressed his mechanical advantage still harder, all while knowing that death hangs on seconds. This is the stark reality he faced: No time to draw attention to keywords, or to describe what is revealing in sentence patterns. No time to understand what lies behind a specific alert, or even determine whether there could be any explanation. The AI's flashing light could only present dispatchers a decision to be made instantaneously, or to be made for them while they hesitated. In that way, cardiac arrest telephone calls resemble one-way airplane tickets and romantic passions and so many of the reasons we want to be alive in the first place: if you stop to ask why, it is already too late.

For his part, Dengel's skin lesion team reminded our group of the General Data Protection Regulation stipulation that data subjects possess a *right to explanation* for any automated decision made by computer algorithms (Lucieri 2020). That will have to change, though, because no one has the responsibility to try the impossible, and the cardiac arrest case demonstrates that under time's pressure and on the edge splitting life from death the only realistic source for trust in AI technology is power as demonstrated by performance. The machine is faster than humans. Which means that the proposal of a *right* to an explanation for an algorithmically generated decision is cancelled by the demands of life itself – not just that patients stay alive, but that life with AI is worth living.

Still, the question remains: Where does the line get drawn? In which instances should trust be built on humanized knowledge about AI decisions? And, when should trust be sought through algorithmic power? More cases will need to be studied, but what these two indicate is that when the moment is critical and the risk is high, power is better than knowledge.

4. Trustworthy or Reliable?

Trustworthy AI is the titular goal of European Commission publications on AI ethics, but should it be? Joanna Bryson (2018) and Mark Ryan (2020) advocate for a shift toward inanimate reliability, and that transition gains support from ground-up work, from investigation that begins with sincere psychological attitudes and bare human experiences.

This is bare experience: trust is inseparable from betrayal. If there was no dishonesty or infidelity than we would not need the concept of trustworthiness. We would not even have it. We could not have it. As Derrida (1998) has demonstrated, these kinds of dialectic word pairings are not just opposites or contrasts, they are episodes of co-dependence: each requires the other in order to produce linguistic meaning. The simplest example may be honesty and lies, if no one ever told the truth, it would be impossible to invent the word or even the idea of lying. This paradox explains the peculiar linguistic experience called bullshitting: it is words, claims, and experiences that are neither truths nor lies, just absurdities.

In lived experience, the fundamental distinction is not between trust on one side and betrayal on the other. Instead, it is between a reality of trust entwined with betrayal on one side, and other realities without either one. As machines are enveloped in language, they too are subjected to the distinction: both or neither.

It is easy to write that trustworthiness is attractive and that deceit is repellent, but judging from how we live, any neutral observer would conclude the opposite: deception and dishonesty are alluring. The evidence is everywhere, but most immediately in our movies and literature, in the places where we *choose* to spend our time. We are drawn to infidelity while scrolling Netflix, we seek deceit while standing in front of the bookracks at the airport. And if these quotidian examples are too crass, then there is the arousal Shakespeare summons from his audience as Brutus plunges his dagger into Caesar's back. It is not a psychotic delight in blood, but the thrill of betrayal captured in Caesar's recognition that it was Brutus too, not just callous assassins driven impersonally by thirst for power. All of this, finally, is *inseparable* from trustworthiness, it is the way the word and concept gain meaning. And all those who venerate the trustworthy are equally engaged by the dark complements, whether they admit it or not.

It follows that if machines are going to be trusted, if we are going to talk and write about them that way, there is a requirement – a condition of the possibility of *being* trustworthy – that they also betray. If the machine wins our trust when it works well, then the machine's failures are not sites of error so much as scenes of unfaithfulness. The AI that falsely signals a cardiac arrest to an emergency dispatcher is not wrong, it is *duplicitous*, and the result should not be disappointment, but *guilt.*

That never happens, though. Not even remotely. In our group's extensive work with two very different startups trying to promote trustworthy AI, not once did a single ethicist, engineer, lawyer, doctor, or manager define false outputs as dishonesty, infidelity, deception, lies, betrayal. No one invoked the idea of being *tempted* by a false positive. It occurred to no one to propose that the light on the emergency call dashboard was winking, trying to lure the dispatcher's attention and seduce with insincere promises. When the machine was wrong, it was just wrong, that was all.

Ultimately, the problem with trustworthy AI, and the reason the idea should be abandoned for the neutral and inanimate term of reliability, is not that we cannot force ourselves as humans to trust the machines. Probably, we can. The problem is what waits on the other side of that trust: mechanical duplicity. So, even if we grant that algorithms *could* be trustworthy, a complete understanding of what that means requires that we also acknowledge engaging with the acidic joys of betrayal. For the human experience of encountering AI today, that joy is inconceivable.

Finally, and stated positively, the way we speak in the real world when machines fail *dictates* the way we must respond when they succeed. With failure, we feel

disappointment, and we talk about error and incorrect outputs. We speak about the opposite of reliability. With success, consequently the decision is already made: AI can be reliable, but not trustworthy.

5. Protection or Performance?

A sentence with jarring implications appears near the end of our group's report on the Copenhagen emergency call case:

> Under the forthcoming Medical Device Regulation in the EU, the AI system will be classified as medical device, and it would therefore need a EC-certification [Redacted].

Blomberg would not have been able to initiate his experiment today. Medically certifying the AI would swamp development in ethical safeguards, including the General Data Protection Regulation, and for good reason: the telephone calls are agony. Loved ones wheezing and clawing with reddening eyes rolling up in their head. Something needs to be done. No one knows what. Gawking onlookers deepen the helplessness. In the tortured environment of cardiac arrest, some people will lose themselves. Others will be themselves, but lose the capacity to control how to exhibit their own identity. Either way, regulatory protections promulgated for personal information in the ethical region of human dignity were crafted for exactly these moments. This is when privacy matters. If tragic emergency calls are not shielded from the commodification of machine learning, it is difficult to imagine what human experience could possibly be considered protected.

On the other hand, commodification works. The machine recognizes cardiac arrest. Lives can be saved.

The dilemma is bottomless, and another example of why people claiming to possess absolute answers to true ethical questions are doing it wrong. Doing ethics right means reaching the point where it is simultaneously true that it is impossible to decide, and a decision must be made. Blomberg was there, faced the impossibility, and decided. The AI development proceeded.

In 2019, Tesla disabled major components of its Autopilot feature in nations where UN/ECE r79 vehicle safety regulations were promulgated (Lambert 2019). The restrictions embodied the European commitment to safeguard against artificial intelligence harms and risks, even at the cost of development and application (Roberts 2021: 1). They also foreshadowed the regulatory paradox that Blomberg initially

escaped, but now cannot. More regulation equals less regulating because innovation is stymied until, theoretically, there is nothing left to restrict.

The practical reality is more complicated. It is always possible for Blomberg and his team to apply for certification from the relevant administrators, manage the bureaucrats, and so continue their work. Still, Blomberg's constriction reveals a worrisome internal logic: it is not just that safeguards erected around human dignity and privacy in the face of oncoming technology are parasitic in the sense that ethicists and regulators need technical advance to provide material for work and justification for existence. There is also the parasitism that, by nature, kills its host.

It does not need to be that way, AI ethics can be reconceived to catalyze innovation. The direct strategy is to credit ethically – not just technically and economically – AI that performs well. Performance *is* a value. One of the most frustrating aspects of our group's interactions with Dengel's skin lesion team and Blomberg's cardiac arrest group was that we found no way to unambiguously account for their pure engineering accomplishments. Instrumental value was easy to locate: both technologies are worth having for the indirect reason that they save lives. But there is something more than that. The skin lesion explainability wrapper and the cardiac arrest natural language detector each stand on their own – without regard for their human benefits – as small but identifiable triumphs of design. They are worth doing intrinsically.

Weighing the value of innovation in a vacuum means, as an extreme example, ethically crediting Nick Bostrom's office product horror story, the one where an AI is programmed to make paperclips and does so, relentlessly (Bostrom 2003). Eventually, the world's natural resources are depleted, and human beings reduced to slavery by the smart machine bending all resistance into the assigned goal of clip production. While that would be a bad end for humanity, it would also be a majestic accomplishment, even awesome, with all that word implies. The best descriptor may be Kant's sense of the sublime, the feeling of reason's power and superiority over nature (Kant 1987: §28). In this way, AI performance as a value resembles art: a dimension of its existence transcends whatever particular effects the work produces for one or another audience.

It is very difficult to find and credit that transcendence for AI ethicists, and especially for those working within the EC *Ethics Guidelines for Trustworthy AI*. The publication does include the value of accuracy, but it is buried deep in the document, underneath the four pillars and then within the category of Technical Robustness and Safety (HLEG 2019: 17). Defined dryly as the "ability to make correct judgements," the idea of simple accuracy is as narrow as it is desiccated. Performance is expansive. In the

cardiac arrest case, speed was as critical as accuracy. In the skin lesion case, the output's elegance was more encouraging than raw calculations of correct predictions. AI that performs well is not only correct and accurate but also quick and graceful.

The four traits of performance as a value in AI ethics are:

- Intrinsically valuable: The only justification required is engineering excellence.
- Independent: The technology is evaluated without regard for its effects on humans, or the world.
- Expansive: The ingredients of performance – what counts as success – cannot be known, defined, or checklisted beforehand, the dimensions of accomplishment are only appreciated after culmination.
- Inevitable: Along with autonomy, dignity, and social wellbeing, accounting for pure performance becomes a required aspect of an ethics evaluation.

The central significance of performance as a value is that it transforms the role of AI ethics from mainstream approaches as represented by the EC *Ethics Guidelines*. The *Guidelines* place the burden on developers to show that their products are trustworthy, which implies fulfilling requirements, including those established to protect personally identifying information. Only then will innovations like Blomberg's emergency call AI be approved for work out in the world. The addition of Performance – accuracy, speed, elegance – to the first line of AI ethics can reverse the priorities. If a machine functions sufficiently well, then it is the regulators who carry the burden of showing why the technology should be constrained. Because creative engineering is understood as *intrinsically* good and worth pursuing, the burden for justifying restrictions falls toward the restrictors.

Instead of the creators proving themselves to regulators, now it is regulators who must prove to creators.

In some cases, the proving will be easy and the addition of performance to AI ethics imperceptible. In Bostrom's paperclip AI the ethics of individual human autonomy and collective social welfare easily overcome the value of the engineering accomplishment. Just because the machine works does not mean it should be built. In close calls, however, in cases resembling the cardiac arrest AI where real human dangers weigh against significant benefits, the accomplishment of the machine itself may prove decisive. If there is no way to be certain beforehand whether an AI ultimately helps or harms humanity, and if the technology performs, then that value endorses and

potentially justifies pushing ahead, through the unknown. Regardless, what is certain is that the EC *Guidelines* do not sanction speculative explorations.

Ultimately, there are two conceptions of AI ethics, one without and the other with performance as a value. One postures defensively against risks. The other leans forward to catalyze opportunities. There will be no way to know which is preferable until it is too late, just as there are no obviously right or wrong places to draw the line between AI ethics as protecting humanity from innovation, and AI ethics as stimulating more of it. But there are different places.

6. Conclusion

Plato taught that ethics and philosophy was about purification: knowledge became truer as it increased in abstraction and decreased in humanity. The light-headedness of intellectual exploration was better than the buzz of wine, the desire for mathematics and science was more passionate and satisfying then sex. That was Plato. The premise of this paper is that Platonic urges have limited AI ethics. Overbalancing principles against practice forces us to believe that we only need to get the abstract, sterile theory right, and then application in the slippery world will come as an afterthought.

These pages are an exercise in reversing the thinking: instead of learning about the world by escaping upward from it, they dive down into it by working with real AI developers as they crash head-on into humanist difficulties. Readers will draw varied conclusions about explainability and performance, and about trustworthiness and reliability, and about the intrinsic value of creative innovation, but the underlying assertion is that those conclusions will emerge stronger from the ground of lived human experiences than from marching algorithms or cloistered philosophy.

Declaration

Author declares that there is no conflict of interest or external funding.

References (pending formatting & inclusion/exclusion from revisions)

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., & Delfino, M. (1998). Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a new 7 point Checklist Based on Pattern

Analysis. Archives of Dermatology, 134(12), 1563–1570. https://doi.org/10.1001/archderm.134.12.1563

Blomberg, S. N., Christensen, H. C., Lippert, F., Ersbøll, A. K., Torp-Petersen, C., Sayre, M. R., Kudenchuk, P. J., & Folke, F. (2021). Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Network Open*, *4*(1), e2032320. https://doi.org/10.1001/jamanetworkopen.2020.32320

Blomberg, S. N., Folke, F., Ersbøll, A. K., Christensen, H. C., Torp-Pedersen, C., Sayre, M. R., Counts, C. R., & Lippert, F. K. (2019). Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation, 138, 322–329. https://doi.org/10.1016/j.resuscitation.2019.01.015

Bostrom, Nick. 2003. Ethical Issues in Advanced Artificial Intelligence Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, pp. 12-17

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., Utikal, J. S., von Kalle, C., Ludwig-Peitsch, W., Sirokay, J., Heinzerling, L., Albrecht, M., Baratella, K., Bischof, L., Chorti, E., … Schrüfer, P. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer, 113, 47–54. https://doi.org/10.1016/j.ejca.2019.04.001

Brown, Tom, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmen. 2017 Adversarial Patch. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://arxiv.org/pdf/1712.09665.pdf

[Redacted]

Bryson, Joanna. 2018.No One Should Trust AI. AI & Global Governance, United Nations University Centre for Policy Research. 2018•11•13 https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html

Chen, Hugh, Lundberg Scott., Lee Su-In. (2021) Explaining Models by Propagating Shapley Values of Local Components. In: Shaban-Nejad A., Michalowski M., Buckeridge D.L. (eds) Explainable AI in Healthcare and Medicine. Studies in Computational Intelligence, vol 914. Springer, Cham. https://doi.org/10.1007/978-3-030-53352-6_24

Derrida, Jacques. (1998). *Of Grammatology*. Baltimore: Johns Hopkins University Press,

Havtorn, J. D., Latko, J., Edin, J., Borgholt, L., Maaløe, L., Belgrano, L., Jacobsen, N. F., Sdun, R., & Agi (2020). MultiQT: Multimodal Learning for Real-Time Question Tracking in Speech. ArXiv:2005.00812 [Cs, Eess]. http://arxiv.org/abs/2005.00812

(HLEG 2019) High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI [Text]. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Kant, Immanuel. 1987. *Critique of Judgment*, trans. Werner Pluhar. Indianapolis: Hackett.

Lambert, Fred 2019 Tesla nerfs Autopilot in Europe due to new regulations, Electrek. May 17. https://electrek.co/2019/05/17/tesla-nerfs-autopilot-europe-regulations/ < https://www.cnet.com/roadshow/news/tesla-model-s-model-x-autopilot-europe-regulations/> https://electrek.co/2019/05/17/tesla-nerfs-autopilot-europe-regulations/

Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., Dengel, A., & Ahmed, S. (2020). On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors. 2020 International Joint Conference on Neural Networks, IJCNN, 1–10. https://doi.org/10.1109/IJCNN48605.2020.9206946

Maaløe, L., Fraccaro, M., Liévin, V., & Winther, O. (2019). BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling. Advances in Neural Information Processing Systems, 32, 6551–6562.

Murphy, D. J., Burrows, D., Santilli, S., Kemp, A. W., Tenner, S., Kreling, B., & Teno, J. (1994). The Influence of the Probability of Survival on Patients' Preferences Regarding Cardiopulmonary Resuscitation. New England Journal of Medicine, 330(8), 545–549. https://doi.org/10.1056/NEJM199402243300807

Morley, J., Floridi, L., Kinsey, L. *et al.* From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. (2020). *Sci Eng Ethics* **26,** 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Roberts, Huw and Cowls, Josh and Hine, Emmie and Morley, Jessica and Taddeo, Mariarosaria and Wang, Vincent and Floridi, Luciano, Governing Artificial Intelligence in China and the European Union: Comparing Aims and Promoting Ethical Outcomes (March 1, 2021). Available at SSRN: https://ssrn.com/abstract=3811034

Ryan, Mark. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Sci Eng Ethics* **26,** 2749–2767 (2020). https://doi.org/10.1007/s11948-020-00228-y

[Redacted]

[Redacted]

[Redacted]

End