by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

What we did

In April 2019, the High-Level Expert Group on Artificial Intelligence (AI) nominated by the EU Commission released the "Ethics Guidelines for Trustworthy Artificial Intelligence", followed in June 2019 by a second document, "Policy and investment recommendations". Initially our group was going to provide a formal response to the pilot phase of the April guidelines, however, as the pilot is geared more towards the business sector, our group undertook a critical assessment of the guidelines from a multidisciplinary perspective.

The Guidelines are aimed at supporting the development of 'trustworthy' AI, by: 1. outlining three characteristics (lawful, ethical, and robust); and, 2. providing seven key requirements (Human agency and oversight; Technical Robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental well-being; and Accountability).

The Guidelines are a significant contribution to the growing international debate over the regulation of AI. Firstly, they aspire to set a universal standard of care for the development of AI in the future. Secondly, they have been developed within a group of experts nominated by a regulatory body, and therefore will shape the normative approach in the EU regulation of AI and in its interaction with foreign countries. As the General Data Protection Regulation has shown, the effect of this normative activity goes way past the European Union territory.

One of the most debated aspects of the Guidelines was the need to find an objective methodology to evaluate conformity with the key requirements. For this purpose, the Expert Group drafted an "assessment checklist" in the last part of the

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

document: the list is supposed to be incorporated into existing practices, as a way for technology developers to consider relevant ethical issues and create more "trustworthy" AI.

In what follows we attempt to assess the implications and limitations of the assessment checklist for the global development of 'trustworthy' AI.

Prologue

During the 1969 Apollo mission to the moon, checklists played such a pivotal role in the logistical operations that astronaut Michael Collins referred to them as "The Fourth Crew Member." As the Apollo mission highlights, checklists are helpful for verifying the presence or absence of factors in many decision-making systems. Since checklists are effective tools for narrowing attention to essential information and eliminating distracting noise, the EU's "Ethical Guidelines for Trustworthy Artificial Intelligence," like many evaluative frameworks, relies on them. Unfortunately, such guidelines may fall victim to the fallacy of the checklist: by over-reductively framing complex problems, they can misleadingly suggest simple solutions to complex, indeed impossible, dilemmas. This poses the danger of creating a compliance regime that allows, even actively encourages, "ethics washing". Technology companies, and companies that use AI in their operations, can be incentivized to minimize their legal liability by touting their conformity to the inadequate guidelines, leaving the rest of society to pay the price for policy-makers not endorsing more nuanced tools.

The EU's Checklist Approach to AI Trustworthiness

The European Commission appointed a High-Level Expert Group on Artificial

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

Intelligence, composed of representatives from academia, civil society, and industry, to support the implementation of the European Strategy on AI. The Expert Group's main remit was to propose recommendations for future policy development on ethical, legal, and societal issues related to AI. The outcome was the development of the Guidelines, with the Implementation List, or checklist, at their heart.

The Guidelines are supposed to contribute to a framework for achieving Trustworthy AI across application domains. The group identifies three components to trustworthy AI: it should be lawful, ethical and technically and socially 'robust'. The Guidelines focus on the ethical component, by first identifying the ethical principles and related values that must be respected in the development, deployment and use of AI systems. The Guidelines aim to operationalize the principles, by providing concrete (although non-exhaustive) guidance to developers and deployers on how to create more "Trustworthy" AI.

The justification of this overall structure stems from the idea that AI is a radically new technology – accurate perhaps in reference to its widespread application, if not to its intellectual lineage – which prompts new regulatory considerations. In this light, a step-by-step process starting with preliminary assessment by means of a checklist can reduce complexity, provide reassurance for developers and deployers, and so help develop trust.

The guidelines mention multiple caveats regarding the practicability, and – of particular importance – what is indicative and what is prescriptive in the structure of the assessment. What also becomes evident upon a closer examination, is that the Assessment List is a tool that may create a false sense of confidence, with little concrete guidance. Indeed, it may be seen as an outdated tool that might work in a cockpit, but in an unpredictable field like AI could end up failing.

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

Technical challenges to AI-related checklists

A key problem for the checklist is that it expects, or indeed requires, concrete answers to technical questions that are so nuanced, that any answer given is necessarily partial. The partial nature of the available answers leads to a certain amount of unavoidable uncertainty, which the checklist does not explain how to navigate. Yet, navigating this uncertainty is in fact the area where guidance is required in order to establish trustworthy AI. We elaborate on this dilemma with two examples: explanation and bias avoidance.

The guidelines ask practitioners to "ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand." Currently, there is no way to definitively ensure that uncontestable explanations exist for a model. Generating explanations is an area of active research, and currently there is no consensus among experts as to what qualities make a good explanation, or even what an explanation consists of. Different methods of generating explanations provide different guarantees, they are often incomparable to each other (and can disagree) because they are based on different baseline assumptions. Thus, they are all partial explanations, with different strong and weak points. Further, the very notion of interpretability is often at odds with values such as security and privacy of information, which can lead to tensions in checking off boxes for both of these desired qualities.

Suppose an individual interacting with an AI system received an adverse outcome and wants an explanation for why the model exhibited this behavior. An example of this may be a resume scanner for job recruitment rejecting an individual from recruitment possibilities for a certain company.

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

One form of explanations are simply alternate inputs that have contrasting results – i.e. other resumes that were accepted by the given company's resume scanner. The individual can then intuit why the model made a given choice based on the differences they perceive in the two resumes. This type of explanation is highly intuitive and easy to understand, but it has its problems. One such problem is that there is no guarantee that any given difference perceived between the two resumes is in fact the model's reason for approving one and not the other. For example, it could be that the resume rejected was from a female candidate and the alternate input provided for explanation was from a male candidate. This may be the biggest difference perceived by the applicant, but it could be that the model is in fact not using this feature at all, but instead the change in hiring outcome was based on a tiny difference in work experience, if the model weights that very heavily. Thus, while explanations are intuitive to understand – it simply involves comparing two inputs – it does not guarantee that the correct reason for the difference in outcome between the two inputs is obvious from this comparison. Secondly, in the interest of privacy, which is another pillar of trustworthy AI according to the EU guidelines, one may want to not give a *real* alternate input for comparison but a fabricated one. Depending on how this alternate input is fabricated, it may be out of distribution of the hiring model's training set, making the model's response to this fabricated input inaccurate, making the entire explanation unreliable.

Another form of explanations are feature-based explanations, which seek to highlight which features (i.e. parts of an input – years of work experience, age, name on a resume) contributed the most to a model's output. In a linear model, this is straightforward—each feature has a weight associated with it and so the pathway from input feature to model decision is clear. However, releasing this information is a security risk that needs to be taken into account when using a linear model.

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

Further, many applications, such as those involving images, require much more complex models. With these more complex models, feature-based explanations become more complicated, as even the question of what *are* the features becomes unclear, due to the fact that the model itself creates its own features as a means of understanding the input as an intermediate step in computation. Some feature-based explanations, in the interest of having more interpretable results, attribute importance to features that may not actually be used by the model. This is because the "features" investigated are defined some independent process (e.g. a person segmenting up an image) that is not necessarily the process that the model uses. [Simoyan, Selvaraju] Conversely, some more cautious forms of explanations potentially miss features that are used by the model.

Beyond this, depending on which baseline assumptions are used in determining which features are most important [Dhamdere, Kindermans, Sundararajan], different feature-based explanations may lead to different results, even if they agree on which features are to be used. It is unclear which are better or more reliable than another. Each method available currently has its own redeeming qualities, but also its own blind spots – and it is often confusing to use more than one type of explanation together because they can provide contrasting results.

Further, all types of explanations can be interpreted widely: there is no one way to interpret the common examples given or the influences of the features returned by a certain method. Even given the same explanation, researchers can come to contrasting conclusions about the behavior of a model. Thus, any explanation necessarily contains some uncertainty.

Navigating these nuanced tradeoffs–intuitiveness versus accuracy of an explanation, and privacy/security versus interpretability—are key places where guidance is

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

required, but not supplied by the checklist. In order to make use of the methods currently available and navigate the inherent uncertainty of the answers they provide, we need to develop answers for questions such as: What purpose is the explanation supposed to serve? Is it for those interacting with the AI system? Is it so that they can understand what they can change about the input in order to get a different result? Is it so that they can be assured there was no mistake? Or unfair bias? Different explanation methods are suited to different applications. Once the purpose(s) are decided, what are the criteria for the explanation to be deemed satisfactory? At what point, and in what applications does a lack of explainability prohibit use of AI in a given context?

Another example of uncertainty comes in bias prevention, when the guidelines ask the developers to "ensure a working definition of 'fairness' that you apply in your AI systems". There are several definitions of what it means for an AI system to be free from bias, or fair [Hardt, Dwork]. While it is necessary to put these processes in place, they will have holes. Different auditing systems and notions of fairness, again, rely on different assumptions, often cannot be used together, [Kleinberg], and each have blind spots. Group-based fairness metrics, such as demographic parity and equalized odds, which seek to treat demographic groups similarly in aggregate (e.g., job recruitment rates should be equal across gender) sometimes sacrifice individual fairness, which seeks to treat similar people similarly on an individual level.

Additionally, these methods of checking for fairness do not often give obvious warning signs. Instead, algorithms often differ from each other in slight gradations, since many "baseline" or "fair" algorithms still have far from perfect scores on these fairness-checking tools. As such, it can be entirely unclear, even to an expert, when to flag an algorithm for exhibiting unfair behavior, outside of blatantly obvious cases.

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

Even if we use all compatible fairness notions together, there are still types of discrimination that we don't have good catches for—such as certain types of smaller subgroup discrimination. So no matter what kind of fairness definition a developer uses, whether they have caught discrimination in the model is uncertain. We need to have some way to *deal with this built-in uncertainty*, including defining the contexts where this uncertainty *prohibits development*.

Experts of the Future and Unacknowledged Uncertainty

The EU assessment list exists among other attempts to manage emergent technologies, such as foresight guidelines for nanotechnology. These are to be used by what Rose and Abi-Rached (2014) have called 'experts of the future', who 'imagine…possible futures in different ways, seeking to bring some aspects about and to avoid others. … In the face of such futures, authorities have the obligation not merely to "govern the present" but also to "govern the future" (p. 14). One way of governing the future is to build tools to manage the creation of new technologies, such as a checklist. Such tools, however, can be ill-suited for technologies where whether the box should be checked or not is uncertain, whether these uncertainties are caused by technological, social, or economic factors. The guidelines aim to provide shortcuts for managing complexity, but the shortcuts are inadequate given the affordances, uncertainty, and pace of AI development.

The Expert Panel acknowledges that the checklist is not a purely mechanistic enterprise or an exhaustive list. Yet in our view, the aspiration that the checklist integrates best practices in AI development and informs the debate unavoidably gives it a powerful orienting force, caveats about its "theoretical nature"

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

notwithstanding. Alternatively, we could begin to think about the checklist in terms of its effect on people, behavior, and – even more importantly – judicial scrutiny of enterprises' activities in developing and commercializing emerging technologies. The checklist was a tool for the last industrial revolution, not the complexity, speed, and uncertainty of the next one.

Ethics on the Ground

The document overlooks the potential disjunction between idealistic ethical guidelines and their practical implementation – ethics on the ground. This echoes the gap between ambitious regulatory visions and the check-the-box corporate compliance that privacy scholars Deidre Mulligan and Kenneth Bamberger documented in their seminal book, *Privacy on the Ground*. It also elides the profound divide existing between ethics and trustworthiness.

It isn't clear what the consequences of these binary questions should be, other than noting that there could be tensions. The checklist as a technology was developed for handling complexity under stress. Experts provide pilots and astronauts checklists as memory aids. This is not a technology for deciding if you should build something at all or how to go about imagining what can be done ethically. Nor is this a technology for dealing with ambiguity. For example, what do you do if the answer to a question is maybe? How do you decide what to do if you can't guarantee that all users understand a decision? Do you then not develop the product at all? Is it likely that a team, having invested significant resources, will stop on a dime because they can't answer a question? More likely, they will simply pass it to the lawyers to finesse.

Thus, while the tool may add "ethical value" for the people who consider it, it does not impose sufficiently specific and consequential requirements to create

trustworthiness. By virtue of its iterative path towards ethics assessment, the document orients itself toward developers and regulators, not the citizens and users whose lives AI will transform. With so much flexibility in interpreting and implementing the tool, and so much complexity and ambiguity in its target, it is natural that in practice, users and other stakeholders will gravitate to simpler versions of the tool. As a result, the checklist is likely to operate as a legalized template to legitimize questionable AI products and practices. A regulatory structure based on checklists will provide some protection, but protection for who? Users or an IT industry that has a practice of breaking things and people first, apologizing second, switching regulatory jurisdictions when they can, and paying fines if they have to? That is not a recipe for ethics by design or trustworthy AI. It is a recipe for "ethics-washing."

Overall, the report's guiding questions are valuable in identifying some of the most salient issues in AI ethics; they help identify issues and problems that all responsible actors in this space should attend to–issues that would be irresponsible to occlude or downplay. Nonetheless, this specific type of utility, heightened ethical awareness, is not what the Expert Panel claims for their framework. Rather, they characterize the framework as the foundation for creating and deploying trustworthy AI. This is more than semantics. Rather, conflating a trustworthy outcome with a more modest ethically attuned one risks encouraging stakeholders to develop misleading, overly-optimistic expectations which could ultimately lead to the opposite of trust–namely, betrayal.

References and Useful Links

Marvin Russell (2017), The History of Checklist:

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

https://hackernoon.com/happy-national-checklist-day-learn-the-history-and-importance-of-october-30-1935-17d556650b89

Astronaut Checklists, https://cultureinfluences.com/en/process/english-guaranty-or-tool/english-astronaut-checklists/

Matthew Hersch, The Fourth Crewmember (2009), https://www.airspacemag.com/space/the-fourth-crewmember-37046329/

Atul Gwande, The Checklist, New Yorker, (2007) https://www.newyorker.com/magazine/2007/12/10/the-checklist

Daniel Greene, Anne Lauren Hoffman, and Luke Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning* (2018): http://dmgreene.net/wp-content/uploads/2018/09/Greene-Hoffman-Stark-Better-Nicer-Clearer-Fairer-HICSS-Final-Submission.pdf

Thilo Hagendorff, *The Ethics of AI Ethics* (2019), https://arxiv.org/pdf/1903.03425.pdf

Deirdre Mulligan and Kenneth Bamberger, Privacy on the Ground (2015), https://mitpress.mit.edu/books/privacy-ground

Johns Hopkins University Center for Government Excellence Ethics and Algorithms Toolkit (Beta) (2019), https://ethicstoolkit.ai/ and https://drive.google.com/file/d/153f0TT_J4cDlr7LRTVKNIDZzu9JUa8ZI/view

Harvard Cyberlaw Clinic, Principled Artificial Intelligence Project (2019): https://clinic.cyber.harvard.edu/2019/06/07/introducing-the-principled-artificial-int

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

elligence-project/

Paula Boddington, Towards a Code of Ethics for Artificial Intelligence: https://www.amazon.ca/dp/B077GCSKB1/ref=dp-kindle-redirect?_encoding=UTF8&btkr=1

John Kleinberg, Inherent Trade-Offs in Algorithmic Fairness https://www.researchgate.net/publication/330459391_Inherent_Trade-Offs_in_Algorithmic_Fairness

M. Hardt, Equality of Opportunity in Supervised Learning https://ttic.uchicago.edu/~nati/Publications/HardtPriceSrebro2016.pdf

Cynthia Dwork, Fairness Through Awareness https://arxiv.org/abs/1104.3913

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron? CoRR, https://arxiv.org/abs/1805.12233

P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (Un)reliability of saliency methods. ArXiv e-prints, November 2017

Rose, N., & Abi-Rached, J. M. (2013). *Neuro: The new brain sciences and the management of the mind*. Princeton: Princeton University Press.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. ArXiv e-prints, 2017.

Simonyan et al. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps"

Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-

by: Antonio Davola, Emily Black, Kalervo Gulson, Geoffrey Rockwell, Evan Selinger and Elana Zeide

based Localization"