Brussels, 18 January 2019
CDO/MvB

EACB response to the Stakeholders' Consultation on Draft Ethics Guidelines for Trustworthy AI

18 January 2019

The **European Association of Co-operative Banks** (EACB) is the voice of the co-operative banks in Europe. It represents, promotes and defends the common interests of its 28 member institutions and of co-operative banks in general. Co-operative banks form decentralised networks which are subject to banking as well as co-operative legislation. Democracy, transparency and proximity are the three key characteristics of the co-operative banks' business model. With 2,914 locally operating banks and 53,000 outlets co-operative banks are widely represented throughout the enlarged European Union, playing a major role in the financial and economic system. They have a long tradition in serving 209 million customers, mainly consumers, retailers and communities. The co-operative banks in Europe represent 81 million members and 719,000 employees and have a total average market share of about 20%.

For further details, please visit www.eacb.coop

EACB AISBL – Secretariat • Rue de l'Industrie 26-38 • B-1040 Brussels
Tel: (+32 2) 230 11 24 • Fax (+32 2) 230 06 49 • Enterprise 0896.081.149 • lobbying register 4172526951-19
www.eacb.coop • e-mail : secretariat@eacb.coop

Introduction

The European Associations of Co-operative Banks (EACB) welcomes the opportunity to provide its input to the consultation on the draft Ethics Guidelines for Trustworthy AI developed by the Commission's High-Level Expert Group on Artificial Intelligence.

Please note that the format below is a copy of the online form stakeholders have been asked to use in order to participate in the consultation.

Consultation on Draft Ethics Guidelines for Trustworthy AI

Introduction: Rationale and Foresight of the Guidelines

Before commenting on the requested section, we would like to share some comments regarding the 'Glossary' (page iv).

The definition of Artificial Intelligence (AI) in the glossary, which is similar to that used in the AI Definition document, differs from more commonplace definitions of AI. The current definition used in the document focusses on one special part of AI, i.e. 'autonomous systems', which is implemented today e.g. in autonomous vehicles. Of course, 'autonomous systems' are a part of AI, but only one among many others. We think that a better starting point would be a more systematic definition of AI in three steps:

1) AI methods including knowledge representation, natural language processing, pattern recognition, machine learning (incl. artificial neutral networks as a subcategory) or machine reasoning. In particular, machine learning ('ML') is a statistical approach to derive (statistical) classifiers based on available data.
2) Decision-making systems make use of AI, but – in a simplified approach – consist typically of two parts:
   a) a classifier (e.g. a credit scoring system with a score value as output);
   b) a decision rule in the sense 'if then else' to compare a score value against a threshold.
   The decision-making can be implemented 'manually' – e.g. with a credit (policy) manual to be used by a human credit expert – or 'technically' with a programme and/or software code, which is written by a human and is the technical implementation of the human intention.
3) Autonomous systems such as self-driving cars, which react in real time and take actions in the real world, i.e. autonomous systems are decision-making systems (as in 2) with real-time processing. Nonetheless, self-driving cars will follow the traffic code (as pre-defined set of rules). 'Virtual' autonomy is the capability to adapt to changing (real-world) situations in a real-time control loop, but based on (maybe rather complicated) rules predefined by human programmers.

It is important to note that none of the above-mentioned systems has an 'own free will' or can make 'individual' decisions, but is always the result of human intention (written as computer code). It is true that human intention is not necessarily correlated to machine output (for example due to errors in the computer code) which highlights the need for robust implementation and testing in order to create trust. However, as we outline in the example below, responsibility remains on the human side.

The 'Moral Machine' experiment (see: Edmond Awad et al., Nature, 24.10.2018) highlights the question as to how an autonomous car should 'decide' in the case of an unavoidable accident (e.g. protection of a young bicycle driver vs. protection of older pedestrians). The possibility for a machine to go through a computer programme in real time and take an action in a dilemma does not represent any ethical agency of the car. The question is: who is responsible for these pre-programmed actions?

Recently Joanna J. Bryson (see: Joanna J. Bryson, Ethics and Information Technology, Vol. 20, Feb. 16, 2018, pp. 15–26) pointed out this question related to the ethics of human decision-making, but not to the programming of machines [quote]: *'The questions of robot or AI Ethics are difficult to resolve not because of the nature of intelligent technology, but because of the nature of Ethics. As with all normative considerations, AI ethics requires that we decide what "really" matters – our most fundamental priorities.'*

The introduction section points out the role of trust and ethics concerning AI technology, which differs from a more traditional approach of risk assessment of a (new) technology, as it has been used in academic research and practical use for decades.

Therefore, we would like to suggest an alternative approach based on the following three steps:

1) Assessment of the technical, social and (maybe) political risks of the use of AI technology;
2) Communication of the risks and benefits of AI technology to society; and
3) Ethical questions of the use of AI technology by human decision-makers.

Such approach would also make it easier to accept the comments raised by some Expert Group members under the section on critical concerns relating to AI. We appreciate why some Expert Group participants might be against including this section – some of the concerns raised are not necessarily directly related to AI as a technology per se, but rather to the way in which the technology is used, to behavioural economics or political choices. This does not mean they do not merit attention from an ethical perspective, but they may not necessarily be translatable into guidelines for AI developers or deployers, which is what the present document aims to do. They would rather fit into the first of the three categories we outlined above '1) Assessment of the technical, social and (maybe) political risks of the use of AI technology'.

Chapter I: Respecting Fundamental Rights, Principles and Values – Ethical Purpose

Concerning chapter I, we would like to provide the AI High-Level Expert Group (HLEG) with some brief comments for each of the sections of the chapter.

- '1. The EU's Rights-Based Approach to AI Ethics'

It would be beyond the scope of this consultation to elaborate on the philosophical relationship between 'rights' and 'ethics'. However, the term 'AI ethics' is misleading. We think that it should be made clear and mentioned in the document that: (i) there is already legislation/regulation, including the General Data Protection Regulation (GDPR), which relates to AI technology; and (ii) the scope of the AI Ethics Guidelines should be 'on top' of existing legislation/regulation, in the sense of specifying the ethical use of AI technology by human beings.

- '2. From Fundamental Rights to Principles and Values'

It would be beyond the scope of this consultation to elaborate on the philosophical relationship between 'values' and 'ethics'. However, we think that it should be made clear that there is a difference between: (i) the assessment of a technology and the decision of a given society to use it or not (e.g. nuclear power or combustion engines); and (ii) the freedom of will – and freedom of contract – of an individual to make a commercial decision (e.g. buy a product or use a service

as offered by a supplier) in compliance with applicable regulations/legislation and based on transparent information made available to the him or her.

- '3. Fundamental Rights of Human Beings'

As already indicated in the title, this chapter relates to human rights in general (respect for human dignity, freedom of the individual, respect for democracy, justice and the rule of law, equality, non-discrimination and solidarity including the rights of persons belonging to minorities, citizens' rights). Of course, they apply to all human relationships in which AI technology is used by one or all participants. However, there is a fundamental misunderstanding as expressed in 3.3: *'AI systems must also embed a commitment to abide by mandatory laws and regulation, and provide for due process by design, meaning a right to a human-centric appeal, review and/or scrutiny of decisions made by AI systems'*. The underlined wording gives the impression that a system based on AI technology can make individual decisions on its own and consequently has to be treated as a new type of social agent. This is a misunderstanding of AI technology in general, as (see above) AI-based systems are always implementations of human intentions and have no autonomy in the sense of an own free will.

We suggest taking into consideration our angle by making it clear throughout the text that human beings are the 'ethic' agents in our society and not technical AI systems.

- '4. Ethical Principles in the Context of AI and Correlating Values'

We think that this this section should be better formulated, as it illustrates the general misunderstanding of these draft Guidelines. Just to give a few examples:

- *'AI systems can do so by generating prosperity, value creation and wealth maximization and sustainability. At the same time, beneficent AI systems can contribute to wellbeing by seeking achievement of a fair, inclusive and peaceful society, by helping to increase citizens' mental autonomy, with equal distribution of economic, social and political opportunity.'*, page 8; and
- *'Avoiding harm may also be viewed in terms of harm to the environment and animals, thus the development of environmentally friendly AI may be considered part of the principle of avoiding harm.'* page 9.

By reading the two examples, it seems that AI systems are the actors in our current society, whereas all the 'values' mentioned refer to human beings as agents in a society, but not to technology. AI systems are – very simply – pieces of technology used by human agents according to their free will.
We suggest specifying this concept throughout the whole document to avoid any misunderstanding.

The issue of 'explicability' is in principle a very technical discussion about the use of statistical classifiers (whether it be traditional scoring based on statistical distributions or classification based on some 'AI' pattern recognition). Beside the discussion in research about the explicability of Artificial Neural Networks (ANN) – with very intensive ongoing research work – any decision-making is usually based on an 'if then else' programme with a scoring value and a benchmark. Scoring algorithms are always statistical classifiers, which require an understanding of the rules of statistics, including the false positive/false negative problem. It makes little difference for the purposes of the paper whether there is any fundamental difference between a complex rule-based classifier and a simple pattern recognition by ANN.

- '5. Critical concerns raised by AI'

As per our suggestion in our introduction, we appreciate why some Expert Group participants might be against including this section. Some of the concerns raised are not necessarily directly related to AI as a technology per se, but rather to the way in which the technology is used, to behavioural economics or political choices. This does not mean they do not merit attention from an ethical perspective, but they may not necessarily be translatable into guidelines for AI

developers or deployers, which is what the present document aims to do. They would rather fit into the first of the three categories we outlined above '1) Assessment of the technical, social and (maybe) political risks of the use of AI technology'.

From this perspective, there is another issue to consider, in particular under the header of longer term concerns: the scenario whereby only a few very large providers control the market for a given service/product, thereby limiting the 'free' choice of customers/users and the possibility to give truly free consent as per the GDPR. Indeed, we should avoid situations where not giving consent could lead to economic/social exclusion. Additional supervisory scrutiny might be warranted.

Chapter II: Realising Trustworthy AI

Concerning the first point 'Requirements of Trustworthy AI', the focus is changed from a discussion of ethics to a discussion of 'social acceptance' of a (new) technology. Nevertheless, it is a crude mixture of technical requirements (very much standard requirements for any new technology) and misunderstandings about AI technology.

As this chapter continues the general misunderstandings about the (human) use of AI technology, we will limit our comments to two examples:

- '4. Governance of AI Autonomy (Human Oversight)'

The quote *'This also includes the predicament that a user of an AI system, particularly in a work or decision-making environment, is allowed to deviate from a path or decision chosen or recommended by the AI system'* makes it clear that the issue is the use of AI in itself, not 'AI autonomy'. Foreseeable implementations of AI technology will be the intention of a natural person ('programmer') and/or legal entity ('company' such as a bank). In the case of a bank, a credit (policy) manual is required by regulations and risk management and has to be followed without any deviation, whether applied by a human credit expert or a technical system. If a rulebook is mandatory, it does not depend on the technology (paper or bits & bytes) that the rule should be fulfilled.

However, the responsibility is always with the people in charge, irrespective of whether a credit manual is approved or an AI-based scoring system is commissioned.

- '5. Non-Discrimination'

The quote *'An incomplete data set may not reflect the target group it is intended to represent'* illustrates the general misunderstanding of the Guidelines concerning AI technology.

First, this statement holds true for any data set used for a statistical classifier (whether traditional distributions or ANN). That is a well-known problem in statistics, but not specific to AI.

Second, there is also the well-known problem that historical data sets will never be 'complete' or 'final' – one always has to find a compromise to optimise false positive/false negative, which has to be an ex-ante compromise and can never achieved 100% for both requirements.

The discussion about 'fairness' is rooted in the 2016 ProPublica publication: 'Machine Bias – There's software used across the country to predict future criminals. And it's biased against blacks' analysing the COMPAS software, which forecasts the probability that criminals will reoffend in the US. This triggered an avalanche of opponents and supporters of this 'algorithmic' approach. Nevertheless, Krishna Gummadi from the Max Planck Institute for Software Systems (see: www.european-big-data-value-forum.eu/wp-content/uploads/2017/12/Krishna-Gummadi-Max-Planck-Institute-Discrimination-in-Machine-Decision-Making-EBDVF17.pdf) offered the best

summary of the misunderstanding of statistics, as all positions are (partly) right. Both sides used different statistical measures to support their claims of the ethical value of 'fairness'.

Concerning the point '2. Technical and Non-Technical Methods to Achieve Trustworthy AI', it applies the well-known 'continuous optimisation' approach to technical systems based on AI technology. Nevertheless, it remains unclear what kind of governance is in scope: the public acceptance of AI technology in general (with an approach conducted by public authorities) or the economic enhancement of AI systems in particular (with e.g. ongoing improvement of false positive/false negative ratios as done with any statistical classifier). In other words: what does 'Trustworthy AI' mean? The adoption of a technology by society ('trusting' this technology) or, rather, understanding statistics (as discussed by Krishna Gummadi)?

While the technical methods are rather standard for any information technology (and therefore not required), the non-technical methods do not have the required clarity. We will give one example taken from the sub-section 'Education and awareness to foster an ethical mind-set': *'Trustworthy AI requires informed participation of all stakeholders. This necessitates that education plays an important role, both to ensure that knowledge of the potential impact of AI is wide-spread, and to make people aware that they can participate in shaping the societal development.'* Even 'normal' statistics is not understood by the majority of society, even less so when it comes to the false positive/false negative problem e.g. in medical diagnostics (!): the expectation of this requirement is very opaque. It is even more obscure what an 'ethical mind-set' should be, as this is not defined and leaves much room for interpretation.

Chapter III: Assessing Trustworthy AI

The list provided in chapter III follows from chapter II and, consequently, has the same problems and flaws. We will only highlight three examples of such flaws:

- '1. Accountability – Who is accountable if things go wrong?'

This is the very usual question of any technical product or service (i.e. liability).

- '3. Design for all – Is the system equitable in use?'

No technical systems can ever be 'equitable': only the use of a technology by human beings.

- '6. Respect for Privacy – If applicable, is the system GDPR compliant?'

Of course, any computer system processing personal data has to comply with GDPR.

General Comments

The draft Ethics Guidelines could be improved by starting with the concept of 'ethical use of AI technology' and a clear understanding of the relationship between AI, traditional statistical classifiers and the use of technology within decision-making processes by natural persons and/or legal entities.

Moreover and as a general consideration, we think that the Guidelines should recognise that the many different use cases of AI cannot always be subsumed under the same ethics principles. Such principles should apply predominantly, if not exclusively, to AI systems that are sufficiently complex and/or handle sufficiently sensitive areas. This would be analogous to a risk assessment under GDPR, where a full DPIA is not always required. Obviously, the class with lenient requirements should be much larger than the one with strict requirements. For example, in Android 9, the phone's operating system uses an AI system (based on deep learning) to highlight apps that a user might need next. AI systems recommending which song to listen to next or which ATM should be refilled next, should be exempt from most if not all of AI ethics' principles. Enforcing principles like beneficence, non-maleficence, justice, and explicability do not make

sense in the above contexts and could potentially hurt the intellectual property rights of the designers. Only the principle of human autonomy, the possibility to opt-out could be implemented in a meaningful way in the above examples.

Contact:

The EACB trusts that its comments will be taken into account.

For further information or questions on this paper, please contact:

- Ms Marieke van Berkel, Head of Department (marieke.vanberkel@eacb.coop)
- Ms Chiara Dell'Oro, Adviser, Retail Banking and Consumer Policy (chiara.delloro@eacb.coop)