Trust in Artificial Intelligence Analysis of the European Commission proposal for a Regulation of Artificial Intelligence

ANTONIO ESTELLA*

I. INTRODUCTION

According to the European Commission, one of the main objectives of the regulatory framework that this EU institution is currently proposing in the field of Artificial Intelligence is to "increment trust in the use of artificial intelligence."[1] Therefore, this paper explores the issue of trust and AI. The questions that it attempts to answer are the following. Why is trust important? Why is trust important, in particular, in the domain of AI? How does the EU Commission intend to achieve the objective of incrementing trust in the use of AI? Will the proposed regulatory framework achieve its proclaimed end?

To answer these questions, this article proceeds as follows. I shall start by reflecting on the importance that trust has for society (section 2). From there, I will define what is to be understood in this paper by trust (section 3). I shall then review the basis of trust (section 4) and shall make a reference to the main sources of evidence on trust (like, surveys and laboratory experiments), and to some of the results that these sources reveal on interpersonal and institutional trust (section 5). In the next section (section 6), I shall go on to analyse specifically the issue of trust in AI, will refer to the existing evidence on the matter, and will review some of the most recent literature on this topic. In the remaining sections (sections 7 and 8), I will describe and analyse the European Commission's proposal for a regulation of AI, and in particular, the part of that proposal that deals with trust in AI. In the last section of this article, I will wrap up the whole argument of this paper and make some conclusions (section 9). The main argument that

1. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}. Brussels, 21.4.2021 COM(2021) 206 final 2021/0106 (COD).

Indiana Journal of Global Legal Studies Vol. 30 #1 (Winter 2023)

will be developed in this paper is that it is inconsequential to speak of trust in AI systems.

II. THE IMPORTANCE OF TRUST

Trust has been defined by some authors as the "lubricant of society"[2] and by others as "a kind of glue that makes society function."[3] Political scientists, economists, and also lawyers have recently centred their intellectual efforts on trying to understand how trust impacts economic growth, development, democracy, justice, and even interpersonal relationships. One particularly clear expression of this renewed interest in trust is the setting up by the OECD of a High Level Group on the measurement of economic performance and social progress.[4] The Group started working in 2013. This group convened eight workshops during the years 2014 to 2016. The latest one took place in Paris in June 2016 and was titled: "Measuring Trust and Social Capital." The outcome of this workshop was published in 2018, together with the rest of the reports of the other workshops that have been mentioned, under the title "Trust and Social Capital."[5] In this paper, Algan gives ample evidence of how trust is positively correlated with economic growth in general and with economic development in particular.[6] The idea is that the more trustworthy a society is, the more it grows and develops in economic terms. The findings of this paper are important since this is the first time that an international institution like the OECD argues that trust should be a necessary component for the measuring of how the nations of the world grow in economic terms.[7]

---

2. *See generally* JON ELSTER, EXPLAINING SOCIAL BEHAVIOR (2d ed., 2015).

3. *See* SCIENCESPO, Joseph Stiglitz on the Importance of Trust in Economics, https://www.sciencespo.fr/en/news/joseph-stiglitz-about-importance-trust-economy (last visted Jan. 1, 2023).

4. Organization for Economic Co-operation and Development [OECD], *High Level Expert Group on the Measurement of Economic Performance and Social Progress* https://www.oecd.org/statistics/measuring-economic-social-progress/aboutthehigh-levelexpertgroup.htm.

5. *See generally* Yann Algan, *Trust and Social Capital*, *in* FOR GOOD MEASURE: ADVANCING RESEARCH ON WELL-BEING METRICS BEYOND GDP 283 (2018).

6. Algan, *supra* note 4.

7. *Id.*

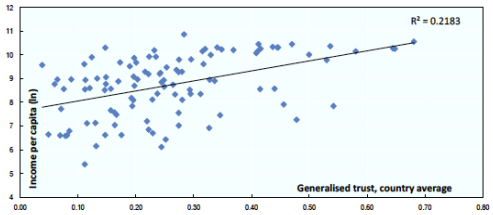


Figure 1: Inter-personal trust and income per capita[8]

On the basis of the previous Figure 1, Algan argues that "countries with higher levels of trust tend to have higher income." For example, Norway has very high levels of trust and has one of the highest incomes per capita of the countries that are included in the previous analysis. An opposite example would be Zimbabwe, with very low levels of inter-personal trust and comparatively low levels of income per capita. Algan acknowledges that there might be problems of reverse causality in analyses on the correlation between trust and economic growth: "one concern has been that this correlation . . . could go the other way around, i.e., from income to trust."[9] However, Algan and other authors have implemented statistical strategies to avoid this effect and try to figure out what direction causality takes in this area: "By focusing on the inherited component of trust, the authors avoid reverse causality. By providing a time-varying measure of trust over long periods, they can control for both the omitted time-invariant factors and other observed time-varying factors such as changes in the economic, political, cultural and social environments." The question is therefore a complex one that needs more refined analyses. However, it is probably safe to say that the positive impact of trust on economic growth and development is undisputed today. Still open to debate and analysis are the specific micro-mechanisms of such correlation.

Similar analyses are being made on trust and democracy, trust and justice, etc. In regards to democracy, the classical reference is Putnam.[10] According to this author, trust is a key component of social capital; therefore, when trust decreases social capital decreases as well, which has a negative impact on democracy. In turn, the impact of trust in the justice system is receiving a lot of attention from different academic

8. *Id.*
9. *Id.* at 302.
10. *See* ROBERT D. PUTNAM, BOWLING ALONE: THE COLLAPSE AND REVIVAL OF AMERICAN COMMUNITY 144–45 (2000); *see* also ROBERT D. PUTNAM ET AL., MAKING DEMOCRACY WORK: CIVIC TRADITIONS IN MODERN ITALY 169–70 (1993).

quarters today.[11] On the one hand, some of these analyses are worrisome of the decline of trust in the judicial system shown by surveys. On the other hand, other analyses have a more positive outlook since, compared to other branches of government, courts seem to be doing better in terms of trust. The debate on the impact of trust on justice and, in general terms, on the legal order is still open; more research still needs to be done in this important area.

III. WHAT IS TRUST?

As we have seen in the previous section, trust is important in different spheres of society, so therefore, we may turn now to the definition of trust.[12] Trust is a very intuitive concept: we all understand what we are talking about when we refer to trust. However, the definition of this concept at a theoretical level is much more elusive. In my opinion, the main reason for this rests in the confusion that exists between trust and cooperation. Trust and cooperation are treated, in many analyses, as co-terminus. However, it is important to differentiate them. Cooperation always stems from interest. I cooperate with you because I have a certain interest in doing it. You cooperate with me because you have a certain interest in doing it. Instead, trust does not necessarily stem from interest. I trust you irrespective of my interest in trusting you. In certain cases, like in blind-trust, trusting someone can even run against my interests. Therefore, the difference between trust and cooperation is that the latter one needs interest, whereas the former one does not.

Starting from this basic differentiation between trust and cooperation, we may use, for example, the definition that the Cambridge English Dictionary gives for trust. According to the CED, trust is "to believe that someone is good and honest and will not harm you, or that something is safe and reliable." I have chosen the CED definition of trust because it puts the focus on one important aspect: that trust is considered as a belief. For the time being, let us restrict the ensuing analysis to interpersonal relationships. When I say that "I trust you," what I am implying is that I believe that you will do what you say you would do. If you say, "I will be back at home at midnight," and I say, "I trust you," what I am saying in just a couple of words is that I believe that you will honour your promise and therefore, you will be back at home at midnight. Trust is therefore a belief, or an expectation, that my

11. *See generally* ANTONIO ESTELLA, LEGAL FOUNDATIONS OF EU ECONOMIC GOVERNANCE (2018).

12. *See generally* RUSSELL HARDIN, TRUST 46 (2006).

interlocutor will respect her commitments.

The idea of commitment is therefore also important for the discussion on trust. As a matter of fact, the clearest way to introduce ourselves in the discussion of trust is to start from a commitment-structure. If you say that "I will be back at midnight," what I am actually doing is making the commitment that I will be back at midnight. Then it follows that if I say "I trust you," what is actually happening is that I am saying that "I trust that you will honour your commitment to arrive home at midnight." We see, therefore, that for trusting structures, we need at least two people. Therefore, we discard situations in which I would say, "I always trusted myself that I would be back at home at midnight." Trust structures only make sense in the framework of commitments, and commitments always involve at least two persons. Therefore, once we have a commitment in place, and trust is at stake, there are two relevant persons: the trustee and the trustor.

Another very common confusion in the domain of trust is to think that trust depends only or mostly on the trusted person. If the trusted person is trustworthy, then we will have all reasons to trust that she will honour her commitments. This is why many analyses on trust posit that in reality, we should speak about trustworthiness instead of speaking of plain trust. I think however that such analyses are misleading, to say the least. The reason is that the trustee and trustworthiness are of course important in a trust structure, but they are not the only relevant players in the game. The trustor is at least as important as the trustee. In particular, the capacity of the trustor to trust plays a fundamental role in this area. To drive the point home, think of the two extremes—and to a certain extent, pathological—cases of blind trust and no trust at all. It would be crazy, for example, to have blind trust in Hitler, as much as it would be odd not to have trust at all in Mahatma Gandhi. The first case would be a case of blind trust and the second case would be a case of pistanthrophobia—the fear to trust anyone. This means that both the trustee's trustworthiness and the trustor's capacity to trust are crucial in a trust structure.

IV. THE BASIS OF TRUST

So, why do we trust (or not)? The bases of trust are also important to consider. For some authors, the bases of trust are plainly rational. I trust you because I have analysed your behaviour and have concluded, on the basis of that evidence, that you are a trustworthy person. Again, it is important not to confuse between rational trust and cooperation and interest. In rational trust, the reason for trusting you is disconnected from my interest. I only trust you because you have said a

thousand times that you would arrive home at midnight, and you have arrived home at midnight.

For other authors, trust has instead a moral component. I trust you because I think it is the right thing to do. In this case, trust is part and parcel of a wider set of principles and beliefs that are constitutive of what we could call the "moral personae." I tend to trust people because my ethics and my morals or my religious beliefs tell me to do so. I tend to trust people because I have a vision of the world in which this would be the right thing to do. Therefore, I do not trust you because I have observed your behaviour and have seen that you tend to honour your commitments, but because I have that moral predisposition to do so. Moral trust plays an important role above all in structures in which the information about the other person is lacking; or, to put it in a different way, in sequential games, in the first move. It is also important to note that rational trust and moral trust are not antagonist concepts: a person holding a moral vision of trust can distrust someone else if she sees that the other one is an untrustworthy person. It is therefore more realistic to think that both types of trust are supplementary. We could say that the person that has a moral vision of trust would need less rational trust and the person that does not hold a moral vision of trust may need more rational trust to trust. The point is here more analytical than normative: to understand why a person trusts, and the extent that she does it, it is important to try to understand what definition of trust she holds.[13]

V. SURVEYS AND LABORATORY EXPERIMENTS

We gather evidence about trust (whether people tend to trust or not, the extent that they do it, how they do it, etc.) through two basic methods: surveys on trust and social experiments. There are a number of surveys that ask about trust, for example, the World Values Survey (WVS).[14] This survey has been asking about trust for at least the last 25 years. In general terms, it can be said that these (and other) surveys differentiate between two basic kinds of trust: interpersonal trust and institutional trust. Interpersonal trust is trust in other people, whereas institutional trust is trust in particular institutions, like the parliament, the government, the political parties, or the judiciary. Additional surveys on trust in AI devices are starting to emerge.

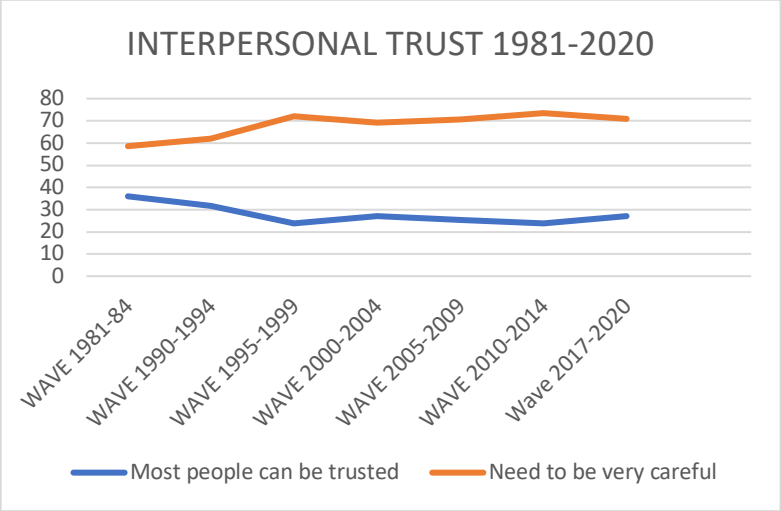
Laboratory experiments are another way to obtain evidence about

13. HARDIN, *supra* note 11.

14. Ronald Inglehart, et al., World Values Survey: All Rounds – Country Pooled Datafile, https://www.worldvaluessurvey.org/WVSContents.jsp (Last visited Jan. 19, 2023).

trust. These are game experiments that help to refine many of the hypotheses that we have about trust. They also give a more realistic, micro-funded, and dynamic picture of how trust structures work in practice. Ideally, surveys should be mixed with laboratory experiments to have a more fine-grained perception of how trust works. My problem with some laboratory experiments on trust is that, in many cases, it confuses trust and cooperation.[15]

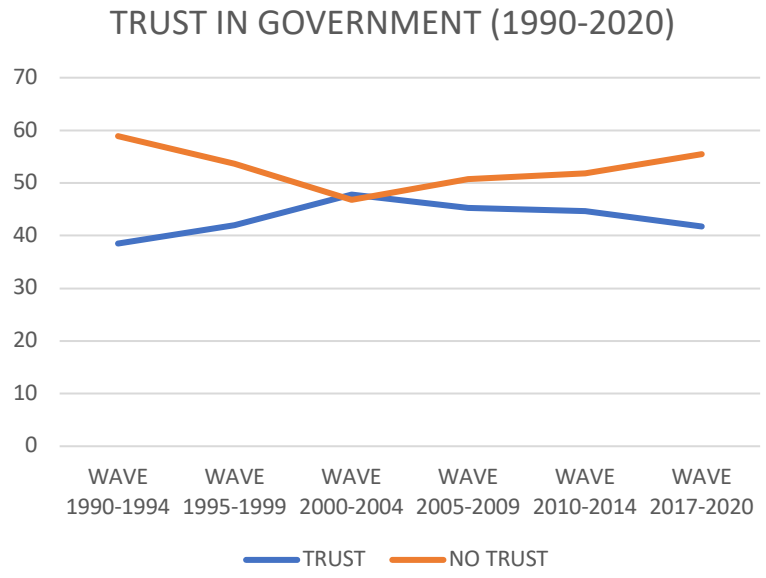
Unfortunately, departure surveys on trust do not provide good news. These surveys show that trust, both interpersonal and institutional, are being depleted all over the globe. Some authors even speak of a "cascade of trust destruction" that could haunt the world.[16] For example, Graph 1 shows the evolution of interpersonal trust from 1981 to 2020. We can easily see that levels of interpersonal trust have not been particularly high across the globe in this time series. We may, however, observe a certain amelioration of this trend since the wave of 2010-2014 to the wave of 2017-2020 (-4 percentage points). However, the difference between those who think that most people can be trusted and those who think that one needs to be very careful is more than 40 percentage points.



Graph 1: Interpersonal Trust (1981-2020)[17]

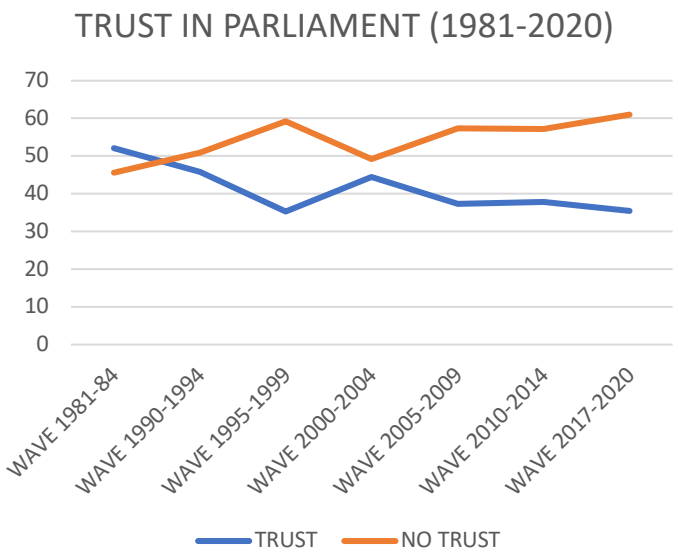
15. HARDIN, *supra* note 11.
16. SCIENCESPO, *supra* note 2.
17. Inglehart, *supra* note 13.

If we turn now to institutional trust, things are not much better either. For example, in Graph 2, the WVS asks people whether they trust their governments or not. The people have been ambivalent across the years, but since the wave of 2000-2004 the trend is clear: mistrust in governments is skyrocketing around the world. The same is true for trust in parliaments. Graph 3 shows that except for the wave of 1981-1984, mistrust in parliaments has been the rule and, once again, since the wave of 2000-2004, it has been growing steadily. The apparent exception to this trend would be the courts. As shown in Graph 4, we observe a reversal of the previous trend of mistrust in this institution after the wave of 1995-1999. Since then, more people seem to trust the courts than not. This finding (why people trust courts and not the other two branches of government: executives and parliaments) is still open for explanation. However, in general terms, we may conclude that both interpersonal and institutional trust are probably at their lowest. This poses problems for our understanding of democracy and for the functioning of the economy, as many analysts have already remarked.

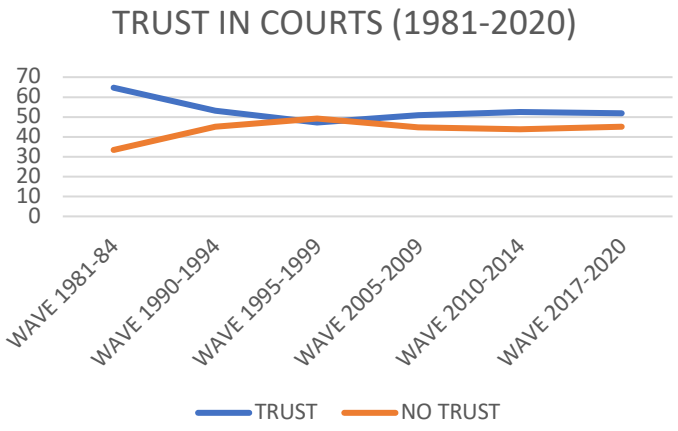


Graph 2: Trust in Government (1990-2020)[18]

18. Inglehart, *supra* note 13.



Graph 3: Trust in Parliament (1981-2020)[19]



Graph 4: Trust in Courts (1981-2020).[20]

19. *Id.*

VI. TRUST IN AI

This is the context in which we should analyse and understand the issue of trust in AI. The question is: if people mistrust other people and the most basic democratic institutions around the globe, why should they trust AI? After all, AI devices are made by humans, not by other machines. Perhaps we should therefore expect that this current wave of mistrust would be replicated in AI.

There are already some (partial) surveys on this matter. All of them point to the same result: in general terms, people tend not to trust AI. In the context of the general mistrust wave that has been previously analysed, this should come as no surprise. For example, Klynveld Peat Marwick Goerdeler (KPMG) conducted a survey on trust and AI in 2021.[21] The surveyed countries were the United States (US), Canada, Germany, United Kingdom (UK), and Australia— five of the most important economies of the world. The outcome of this survey is dismaying for the prospects of AI. In effect, as shown in Figure 2, only 28% of the sample would be willing to trust in AI, the highest being Australia (32%) and the lowest being the UK (26%). This survey also asks about trust in AI healthcare devices, which presents somewhat better results: 37% of the sample would be willing to trust healthcare AI. According to this report, trust, or rather lack thereof, in AI is influenced by four major causes: beliefs in the capacity of the regulatory system to make AI use safe; beliefs in the perceived impact of AI in jobs; familiarity and understanding of AI; and beliefs on the uncertain impact of AI on society. Of the four causes, regulation is clearly the strongest driver. This means that if people believed that the AI regulation in place was adequate, then they would have a better opinion on the other three items. In other words, at least according to this survey, the EU Commission has all reasons to focus on the regulation of AI as a way to enhance trust in AI.

---

20. Inglehart, *supra* note 13.

21. Nicole Gillespie et al., *Trust in Artificial Intelligence: A Five Country Study*, KPMG (March 2021) https://home.kpmg/de/en/home/insights/2021/06/artificial-intelligence-five-country-study.html.

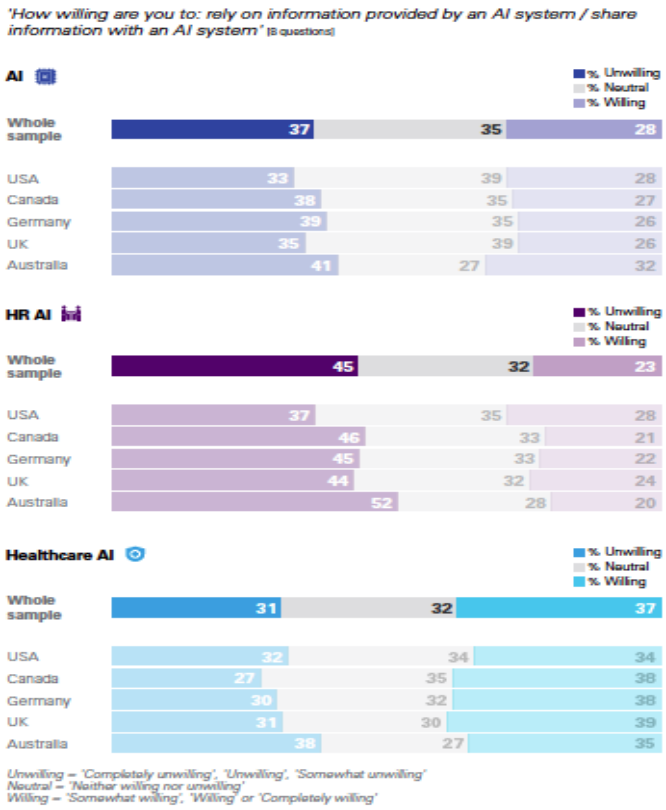


Figure 2: KPMG survey on Trust in AI[22]

In turn, a survey conducted by Ipsos in 2022 nicely complements the previous KPMG survey.[23] According to the Ipsos survey, only 50% of the sample would "trust companies that use artificial intelligence as much as they would trust other companies." Asked about the benefits and drawbacks of using AI, the percentage of people who think the benefits outweigh the drawbacks are the following: UK 38%; Australia 37%; Germany 37%; USA 35%; and Canada 32%. This means that for a vast majority of the people in these five important economies of the world, the drawbacks of using AI devices are much higher than the benefits. It is possible to think these findings correlate with trust in AI.

22. *Id.*

23. *See* Nicolas Boyon, *Global Opinions and Expectations about Artificial Intelligence*, IPSOS (January 5, 2022) https://www.ipsos.com/en/global-opinions-about-ai-january-2022.

Choung, David, and Ross have explored the issue of the impact of trust in AI voice assistants like Siri, Alexa, and Google Assistant. Voice assistants have been designed to be more human-like and therefore more trusted devices.[24] The previous authors develop two studies in their paper. In the first study, they find that trust is key to build positive attitudes towards AI voice assistants. Therefore, "people are inclined to regard a technology beneficial if they trust it." In contrast, "a lack of trust could raise concerns about the potential threats and risks of the technology instead of its benefits." The paper finds that in using AI voice assistants, ease of use and perceived usefulness are better predictors than trust. However, they also point out that trust can influence the other two factors.

In the second study, the impact of the "human-like dimension of trust" and the "functionality dimension of trust" on AI is tested. The human-like dimension of trust is to attribute human characteristics to AI (like the social and cultural values of the algorithms, for example, or human physical characteristics, as in robots). The functional dimension of trust is that AI works properly. According to this second study, the two factors are significant in predicting the perceived usefulness and positive attitude towards smart technology, which in turn predicts greater usage intention.

In turn, Lockey proposes with particular clarity the problems derived from some of the perversions of trust, like blind trust, or blind faith, as they call it.[25] "A foundational tenet of trust theory is that . . . it should be based on "good reasons"; trusting with no good reasons is no trust at all." These authors analyse five challenges that are directly related to trust in AI: 1) transparency and explain-ability; 2) accuracy and reliability; 3) automatism versus augmentation; 4) anthropomorphism and embodiment; and 5) mass data extraction. These authors find that the enhancement of all these factors increases trust, but with some qualifications. For example, over-explaining, in particular contexts like in AI assessment grading tools, may serve to decrease trust. Another finding is that accuracy is not enough. In some contexts, like in large, street-based games, the perception of accuracy may be as important as accuracy itself. Further, the issue of automatization versus augmentation is particularly relevant for AI in healthcare. In a series of experiments that are reported by the authors, it was found that people tend to trust less automated advices in healthcare than augmented ones (that is, advices that are made by AI

24. *See generally* Hyesun Choung et al., *Trust in AI and its Role in the Acceptance of AI Technologies*, INT'L J. OF HUM.-COMPUT. INTERACTION (forthcoming March 2022).

25. *See* Steven Lockey et al., *A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities, and Future Directions*, 54 L. HAW. INT'L CONF. SYS. SCI. 5464–67 (2022).

but supported by a human physician). The "human in the loop" approach is also preferred in the field of financial services. In turn, anthropomorphism is seen by the authors as a double-edge sword; in principle it increments trust, but it can also develop into over-trust. For example, the authors report that a study on an anthropomorphic health-care robot was perceived as less trustworthy than a machine-like robot. Finally, in regards to mass data extraction, the issue of privacy, and the use of data when using AI clearly impacts trust in AI; however, the authors report that more empirical work needs to be done in this area.

Winfield and Jirotka explore, at a more theoretical level, the connection between ethics and trust. According to the authors, a more inclusive, transparent, and agile form of governance would serve to build and maintain public trust in AI and ensure that AI is developed for the common interest. In this connection, these authors make a number of recommendations that range from publishing ethical codes of conduct and providing ethics training for all and being transparent about ethical governance of AI.[26]

In turn, Afroogh highlights, in his analysis on trust and AI, that mistrust in AI is a crucial barrier for its development.[27] According to this author, "any future development, implementation and usage of AI are tightly related to the public trust and supportive stance."[28] He therefore proposes a probabilistic theory of trust, the core of which is the distinction between four kinds of situations: an AI agent's trust in another AI agent; a human agent's trust in an AI agent; an AI agent's trust in a human agent; and an AI agent's trust in an object. His probabilistic theory would be formulated as follows: "A (including a human agent, AI, etc.) trusts B (including a human agent, AI Intelligence, etc.) or A believes that B is trustworthy only if there is a high degree of imprecise probability that B represents the proper functions or competence in nearby possible worlds." According to Afroogh, his formulation would integrate the four kinds of situations that I have mentioned before.

In sum, the previous review of surveys on trust and of the most recent academic literature that deals with this evidence points at two directions: the first one is that in general terms, people tend to distrust AI. The second is that it is possible to think that trust in humans or institutions is probably a different phenomenon from trust in intelligent

---

26. Alan F.T. Winfield & Marina Jirotka, *Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence*, PHIL. TRANS. R. SOC'Y A., Aug. 21, 2018, at 1, 1, 10 (2018).

27. Saleh Afroogh, *A Probabilistic Theory of Trust Concerning Artificial Intelligence: Can Intelligent Robots Trust Humans?*, AI & ETHICS, June 2, 2022, at 1, 1, 13–14.

28. *Id.*

machines. Maybe our theoretical understanding of trust should be further refined or even reformulated to integrate structures in which humans try to trust in AI. Or maybe it is not only a matter of qualification or even reformulation: perhaps when we speak about trust in machines, even if they are intelligent machines, we would be actually thinking in a different situation. For example, Afroogh differentiates between trust and reliance, and questions whether we should speak more of reliance in AI machines than in trust in AI machines. On the other hand, it is clear that we project the idea of trust to agents that are non-human (like institutions), and we think that this is not a contentious issue. Maybe when we say that we trust an institution, what we are implying is that we have trust in the persons that compose that institution. For the same token, maybe when we say that we trust an AI healthcare device, what we are implying is that we trust the humans that fabricated it and that are behind the machine. The whole thing would of course get much more complicated when, and if, AI machines become completely independent from humans.

VII. THE EUROPEAN COMMISSION'S PROPOSALS FOR ENHANCING TRUST IN AI

Concerned with the problems of trust in AI, the European Commission proposed a new regulatory framework that attempts to mitigate the detected problem of mistrust in AI devices. To this end, the European Commission issued a White Paper in February 2020, "On Artificial Intelligence-A European Approach to excellence and Trust."[29] This White Paper was followed by a Commission Communication of March 2021, "Fostering a European Approach to Artificial Intelligence,"[30] which was published together with a proposal for a regulation "laying down harmonised rules in Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts" of March 2021.[31] We shall review these three documents in the

29. See generally Commission White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (Feb. 19, 2022). See also White Paper on Artificial Intelligence: A European Approach, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12270-White-Paper-on-Artificial-Intelligence-a-European-Approach/public-consultation_en (June 14, 2020).

30. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Fostering a European approach to Artificial Intelligence. Brussels, 21.4.2021 COM(2021) 205 final.

31. Commission Proposal, *supra* note 1.

following subsections.

A. The Commission's White Paper on Artificial Intelligence and Trust

The point of departure of the European Commission's White Paper is the assumption that trust is a prerequisite for the uptake of AI. Accordingly, the White Paper presents a number of policy options to enhance trust in AI. One idea is floating over the whole document: this is the notion of creating an "ecosystem of trust." For the creation of an ecosystem of trust in the field of AI, the White Paper tries to identify the main risks that may yield problems of trust in the domain of artificial intelligence. We shall see later on that the whole European Commission's regulatory framework in this field pivots around the idea of risks for trust. To deal with these risks, the Commission proposes to adapt the existing EU legislation on product safety and liability to the requirements of building trust in AI. A second proposal is to adopt a specific regulatory framework in the field of AI.

The White Paper was open for public consultation and comments since its publication in February 2020 until May 2020. The result of this public consultation was 1216 comments, which mainly came from citizens (30%), undertakings (18%), and academic institutions (12%). The rest of the comments had a diverse origin (entrepreneurial associations, NGOs, public administrations, and extra-European Union citizens).[32]

B. The European Commission's Communication on AI

In the European Commission's Communication on AI, the Commission announced that it is proposing a regulatory framework on trust in AI, and it also explained the main philosophy behind this new regulatory framework. As said before, the whole European Commission's edifice in this field revolves around the idea of risks. The Commission indicates that there are three kinds of risks: risks that are considered to be unacceptable, and therefore, are banned; high risks, that are to be highly regulated; and other (minor) risks that have a more lenient regulation. For example, the use of AI to contravene the European Union's values and violate its fundamental rights are to be banned. A particular case that the Commission mentions in its communication is that of remote biometric identification systems. An

32. *Commission White Paper*, *supra* note 29.

example would be the real time use of AI for law-enforcement purposes, which would in principle be prohibited, unless when exceptionally authorised by law. This authorisation would be subject to specific safeguards.

In regards to high-risks, the European Commission specifically mentions the example of AI systems intended to be used to recruit people or to evaluate their creditworthiness and also the case of judicial decision making. These high-risks AI systems would not be prohibited but would be subject to the fulfilment of strict requirements and obligations. Finally, regarding the other (minor) risks, the uses of AI would be subject to the compliance of minimal transparency requirements. The European Commission cites, in particular, the examples of chatbots, emotion recognition systems, and deep fakes as examples belonging to the category of "minor risk." The European Commission summarizes its regulatory approach on trust and AI as "enabling trust without preventing innovation."[33]

C. The European Commission's proposal for a regulation on Artificial Intelligence

One of the declared fundamental aims (but of course, not the only one) of the European Commission's proposal for a regulation in the field of AI[34] is to mitigate the problems of mistrust in AI that the Commission, and other stakeholders in this area, have well identified. In this area, and as has been said before, this long proposal for a regulation (more than eighty articles) pivots around the notion of risks associated with the problem of trust in AI. The three categories are unacceptable risks, high risks, and other risks.

Article 5 of the proposal sets up a list of "prohibited AI practices." The idea is, therefore, not to ban specific types of AI but the use of specific types of AI. The prohibited practices are the following:

- The use of AI systems that distorts a person's behaviour in a manner that causes, or is likely to cause, physical or psychological harm to humans.

- The use of AI systems exploiting vulnerabilities of a specific group of persons linked to age and disabilities, so that AI materially distorts the behaviour

33. Communication from the Commission, *supra* note 30.
34. Commission Proposal, *supra* note 1.

> of one person belonging to that group in a way that produces or is likely to produce harm.
>
> - The use of AI systems for the evaluation or classification of the trustworthiness of natural persons, so that the outcome is detrimental for those persons.
>
> - The use of "real time" remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement, unless and as far as such use is strictly necessary for a number of objectives which are related to individual and collective security threats.

Of the four cases, the hardest one is the last case, since it admits exceptions. According to the proposal, the exception is subject to a prior authorisation which shall be granted by either a judicial authority or an independent administrative authority of the Member State in which the use is to take place. It is for the Member States to develop the precise procedural requisites that are to be applied to authorisations, within the limits set by the proposal. These limits are, in essence, the following: the authorising agency has to take into account the seriousness, probability, and scale of the potential harm in the absence of the use of the AI system; and it has to take into account what consequences would be derived from the use of such system in terms of fundamental rights and freedoms.

In turn, Title III of the proposal (articles 6 to 51) is the lion's share of the regulation. It regulates the so-called "high-risks." The structure of this Title is the following. It is divided into five chapters, which deal with the following issues: classification of AI systems as high-risk (Chapter 1); requirements for high-risks AI systems (Chapter 2); obligations of providers and users of high-risk AI systems and other parties (Chapter 3); notifying authorities and notified bodies (Chapter 4); and standards, conformity assessment, certificates and registration (Chapter 5).

Chapter 1 defines what is to be considered as an AI system use that has a high-risk. To be considered as having a high-risk, the AI system use has to fulfil two conditions: first, that the AI system is used as a safety component of a product, or is itself a product, listed in Annex II of the proposal; and secondly, that the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a conformity assessment according to Annex II of the proposal. Further, all the AI systems listed in Annex III of the proposal are

considered to be of high-risk. Therefore, the proposal remits us to Annexes II and III of the regulation. Annex II includes the list of "European Union harmonised legislation based on the new AI Legislative Framework" as well as the list of "other European Union harmonised legislation." This is a list of directives and regulations relating to the field of AI. One should therefore take a look at each and every one of these pieces of legislation to try to make sense of which AI uses are considered high-risk. It would have been more transparent to extract those uses from the previous legislation and incorporate them in an annex. Instead, Annex III includes a proper list of the AI uses that are considered to have a high risk. Some examples are: biometric identification and categorisation of persons, management and operation of critical infrastructure, education and vocational training, and administration of justice and democratic processes. For example, regarding the latter, all AI systems that are intended to assist a judicial authority in researching, interpreting facts, and applying the law to a concrete set of facts would be considered high-risk.[35]

Chapter 2 regulates the requirements for high-risk AI systems. These requirements have to be complied with by users and providers of AI systems (depending on the requirement). Some of these requirements are the following: to establish, implement, document and maintain risk management systems for AI; to set up training, validation and testing of data sets for AI systems; to draw up technical documentation for high-risk AI systems; to design AI systems with the capability of automatic recording of events; to design and develop high-risk AI systems in such a way as to ensure that their operation is sufficiently transparent, to enable users to use the AI systems in a proper way; and to design and develop high risk AI systems in such a way that they are accurate.

Let me draw the readers' attention to the requirement that is established in Article 14 of the proposal. This requirement is human oversight. According to this article, high-risk AI systems have to be designed and developed in such a way that they allow for effective human oversight. The key obligation established in this article is found in letter "e", of paragraph 4 of the commented Article 14. It reads as follows: "[H]umans overseeing a high-risk AI system must] be able to

35. *See* Invertia: El Español, Justice Awards Telefónica a Project that will allow 3,000 Judges to Issue Sentences with Voice and Artificial Intelligence, https://www.elespanol.com/invertia/empresas/tecnologia/20220606/justicia-adjudica-telefonica-permitira-sentencias-inteligencia-artificial/677432697_0.html (Last visited 7 June 2022) (the Spanish online daily "El Español" reports that the Spanish Department of Justice has granted a contract to Telefonica, one of the most important Spanish telecommunications companies, to assist judges to write judicial decisions with the help of Artificial Intelligence. This would be a case that would probably fall in Annex III.)

intervene on the operation of the high-risk AI system or ***interrupt the system through a "stop" button or a similar procedure.***" (emphasis mine). This means that behind any high-risk AI system there must be, at the end of the day, a person. This crucial point has more implications than it initially seems regarding our understanding of the relationships between trust and AI, especially when the AI systems are considered to potentially yield high risks. It means that when we are speaking of trust in AI in reality we are speaking of trust in the person that is behind the AI system. As the proposal for regulation clearly contemplates, the problem is when intelligent machines acquire complete independence from humans. This fundamental point of the whole European Commission's regulatory edifice on AI is discussed in section 8.

In turn, chapter 3 of the proposal establishes a number of obligations upon providers and users of high-risk AI systems. It is difficult to draw a line between requirements and obligations. The logic of the proposal seems to be that a requirement is a characteristic that the high-risk AI system must have, whereas obligations are imposed on persons, natural or legal. However, it is obvious that many requirements imply, be it indirectly, correlative obligations. In any case, the obligations set up by chapter 3 are the following: to ensure that the requirements that have been seen before are complied with; to have a quality management system in place; to draw-up technical documentation of the high-risk AI system; to ensure that the high-risk AI system undergoes the relevant conformity assessment procedure, prior to its marketing; to comply with a number of registration obligations; to take the necessary corrective measures; to inform the national competent authorities of the Member States of the non-compliance and the corrective measures that have been adopted; to affix the CE marking to the high-risk AI systems; and to show the high-risk AI system's conformity with the requirements previously seen.

Once again, the human factor is the key in the domain of these obligations. In effect, article 29(2) says that "the obligations . . . are without prejudice of the user obligations under European Union or national law . . . for the purpose of implementing the human oversight measures indicated by the provider.". Therefore, it is for the provider to set up the "stop button" and for the user to press it if there is a need. On the one hand, this is a very good illustration of what has been indicated before: requirements include implicit obligations. Here, the obligation is for the provider to establish a disconnection system in the high-risk AI system. In turn, the chapter on obligations, chapter 3, specifies that it is for the user to disconnect the intelligent machine.

Chapter 4 makes a difference between notifying authorities and notified bodies. The idea is that competent national authorities, which

the Member States must designate, have to extend accreditations for bodies that evaluate the conformity with the proposal's requirements of high-risk AI systems. The proposal for regulation establishes a clear dividing line between national authorities and conformity bodies: the former must avoid conflict of interests with conformity assessment bodies, and they must ensure the impartiality and objectivity of the operations undertaken by the latter. Here it would have been important to explicitly prohibit any revolving door system between the two.

Finally, chapter 5 regulates the substance of the so-called "conformity assessment" for high-risk AI. As points 63 and 64 of the preamble of the proposal indicate, the idea here is two-fold: first, conformity assessments should be carried out according to the sectoral legislation relating to AI (for example, the Machinery Regulation). The second idea is that to minimize the economic impact of conformity assessments upon AI providers, AI Providers should carry out their own conformity assessments under their own responsibility as a general rule. The main exception is to be found in AI systems intended to be used for remote biometric identification of persons, for which a third-party conformity assessment is made compulsory by the proposal. For these third-party conformity assessments, notified bodies should be designated, as has been seen before. In turn, Annex VI of the proposal specifies what the conformity assessment procedure based on internal control is. In this procedure, the drawing up of a specific technical documentation is the key. The technical documentation to which this annex, and also article 11 of the proposal, refer is established in Annex IV of the proposal. Further, article 48 of the proposal establishes the obligation for the provider to draw up an "EU declaration of conformity." The content of such declaration is established in Annex V of the proposal. Additionally, article 49 establishes that the Conformité Européenne (CE) shall be affixed visibly, legibly, and indelibly for high-risk AI systems.[36] Finally, article 51 specifies that providers of high-risk AI systems of article 6(2) (those listed in Annex III of the proposal) shall register the information established in Annex VIII of the proposal in the EU database that is established in article 60 of the proposal. In sum, the conformity system that is established by the proposal does not differ much from other products' conformity assessment systems, although it has some exceptions.[37] The system is therefore very much based on the individual responsibility of providers. Taking into account the specifics

36. Commission Regulation, Product Requirements: CE Marking https://europa.eu/youreurope/business/productrequirements/labelsmarkings/cemarking/index_en.htm (Nov. 21, 2022) (providing the EU general conformity system described in general terms).

37. *Id.*

of the AI field, this part of the proposal can be open to criticism.

Title IV of the proposal regulates "the other" uses of AI systems; that is, the AI systems' uses that do not belong to the two categories that have been reviewed so far (prohibited uses of AI systems and high-risk AI systems). This Title is composed of only one article, article 52, which simply establishes transparency obligations for providers of AI systems. The first obligation is that AI systems shall be designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obviously the case. The second obligation is that users of an emotion recognition system or a biometric categorisation system must inform of the system's operation to the natural persons that are exposed to them. A third obligation imposed on users as well is a disclosure obligation: users of AI systems that may yield so-called "deep fakes" must disclose that contents have been artificially generated or manipulated. Article 52 establishes a number of exceptions for each of these cases; it also establishes that it cannot affect the requirements of Title III, previously analysed.

VIII. ANALYSIS

The proposal for a regulation of AI tries to address the issue of trust in AI, among other objectives. We have seen in the previous section that the approach taken by the Commission is decremental: the Commission identifies uses of AI systems that generate so many risks that are prohibited; further, it identifies uses of AI systems that generate high risks, so as to need a conformity assessment plus other compliance requirements; and finally, it identifies uses of AI systems that are only made subject to transparency obligations.

As has been suggested in the previous section, the main requirement or obligation that the proposal establishes is that of human oversight. A human has to be behind the intelligent machine. She has to ensure that the intelligent machine is under control. She also has to insert mechanisms in the intelligent machine that allow the human to interrupt its operation, and she has the obligation to report malfunctioning. This is the so-called "human in the loop" approach.

From a theoretical perspective, the fact that AI systems have to rely on the "human in the loop" approach poses some fundamental questions relating to the relationship of trust and AI. In turn, this theoretical perspective has clear practical and legal implications. The theoretical problem that is posed by this approach is that, in reality, it is inconsequential to speak of trust in AI. The practical effect is that if it is not possible to speak of trust in AI systems, then the Commission's attempts to enhance trust through risk configuration will hardly be

successful.

Let me start with the first point— the theoretical point. Trust structures are always, and I would add, only, conceived among human actors. When we speak of trust, we are speaking of an attitude, an emotion, or a rational expectation that can be only proclaimed for humans towards humans. We can speak of trust in humans because we know more or less how human rationality and behaviour work. Therefore, we can elaborate on expectations about humans' behaviour. Precisely the outliers of common rational behaviour are understood by social sciences as deviations from the rule— as pathological states of mind— that need a different treatment. For example, Elster's analysis on addictions[38] is important because addictions escape from our traditional understanding of common rationality. The problem is that we cannot apply this scheme to intelligent machines. We simply do not know how machines are going to behave when they acquire autonomy and independence from humans. We simply do not know how the intelligent machines' rationality (if we can speak of a machine's rationality) is going to evolve in the future. The issue of independence from humans is what is at stake here. Therefore, trust is projected from humans to humans since we know how humans act. However, it is impossible to project trust from humans to intelligent machines since we do not know how intelligent machines act. The matter is, as said before, not only practical, but above all theoretical. Assume that we would understand in the far future AI's rationality, but even in this case, it would not be a human rationality. Therefore, we could not speak of trust in AI systems even if this assumption was ever held.

Maybe a different perspective can help to clarify this theoretical point. This perspective is that of animals. As has been convincingly argued by some authors, animals are "sentient" beings.[39] This means that they are able to feel. This fact has been tested in laboratory experiments with animals and today is beyond any reasonable doubt. From this evidence, many authors have promoted an "animal rights' agenda," that has made headway in some states of the world. Now animals have more protection from humans than was the case a century ago. The point is that asserting that animals have sentiments is to indirectly say that animals have a kind of rationality. In effect, the most recent approaches in rationality argue that the divide between feelings, or emotions, and rationality, is plainly absurd: emotions are part and parcel of our rationality. Being the case, then animals have a certain kind of rationality. This rationality is a sort of human-downgraded

38. JOHN ELSTER, STRONG FEELINGS: EMOTIONS, ADDICTION, AND HUMAN BEHAVIOR 190-91 (Francois Recanati ed. 1999).

39. PETER SINGER, ANIMAL LIBERATION 47, 49 (2d ed. 1995).

rationality. This means that it is close to human rationality but it is, in general terms at least, much less developed than the human rationality. Can we therefore say that I can trust animals? Can we therefore say that I trust my dog, or my horse? The answer to this question would be positive. Yes, we can say that we can trust an animal, because animals have a kind of rationality that is very close to human rationality. It is a kind of rationality that we more or less understand and more or less can control. Of course, this type of trust will be much more nuanced and qualified than trust in humans. But the fact that animal rationality is close to human rationality makes it possible to say that we can trust an animal.

This is not the case for machines. Machine learning is the problem in this domain. According to many analysts, once certain algorithms are in place, some intelligent machines learn in a way that we can simply not understand.[40] And this is not going to stop here: on the contrary, in the future, we are going to have many more instances in which this is going to be the case. If we do not know how machines learn, then we do not know about their rationality since learning is the main gateway to rationality. This means that we should discard any discussion of trust from humans towards AI systems. As Afroogh suggests, we might rather speak of reliance in AI systems instead of trust, for example.[41] In other words, the discussion on trust in AI seems misplaced from a theoretical perspective.

This has, as previously announced, clear practical implications. The main implication is that, as some authors argue, a good dose of mistrust in machines whose rationality we cannot comprehend would be a case in point. Said in other terms, for human rationality, mistrust in AI systems (I insist once again: in systems which rationality we cannot understand) is the appropriate outcome. Therefore, my proposal is to move to a different practical dimension and speak of other concepts that are closer to the world of machines. It is obvious that an AI system is a product different than other products. But from there it is difficult to give human qualities and characteristics, such as trust, to AI systems. In other words, we cannot simply speak of such a humane activity regarding machines as trust is. The idea, from a practical perspective, would therefore be to place this debate in the realm of reliability, accuracy, efficiency, correctness, technical competence, control, etc. In

40. Bradley (2017), writing for Forbes, reports that "Facebook shut down an artificial intelligence engine after developers discovered that the AI had created its own unique language that humans can't understand". See the article here: https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/?sh=139144ec292c

41. Afroogh, *supra* note 26.

other terms, what I want from a machine, even from an intelligent machine, is that it works properly and accurately under my control. Therefore, it would be important to apply a healthy measure of mistrust towards it, above all if it has intellectual properties that may easily escape my control.

IX. CONCLUSIONS

I have analysed in this paper the European Commission's proposal for a regulation of AI systems, in particular the part of this proposal that is aiming at mitigating the problems of human mistrust in AI systems. This analysis cannot be done if it is made outside the context of a wider discussion on trust. This is why I started this paper by making a number or reflections on trust: its meaning, its importance, the attention that it is receiving today from the academic world. I defined trust as an expectation about whether or not my interlocutor will honour her commitments. Trust analyses only make sense within commitment structures, in which A commits to doing X, and then B trusts (or not) that A will honour her compromise. Commitment structures are two-person games, as are trust structures. I also underlined the importance of bearing in mind that both the trustor and the trustee (and not just the trustee) are important to understand how trust works. The trustee has to have trustworthiness, but trust is impossible when the trustor is uncapable of trusting, even in contexts in which the trustee's trustworthiness is Mahatma Gandhi-like.

All this projects a picture in which trust is understood, as a matter of both theory and practice, as a very human behaviour. This is why this paper's main argument is that it is odd to speak of trust in the context of machines, even if they are intelligent. Expectations of trust are based on a given common knowledge about how rationality works and what rationality is. This is something that we cannot and will not be able to predicate about intelligent machines.

The Commission's approach to solve the problems derived from mistrust in AI systems is based on the conception that the regulation of risks impacting trust will be enough to at least mitigate the current wave of mistrust that society has towards AI systems. Once these risks are properly regulated, then the outcome should be an enhancement of trust in intelligent machines. Therefore, the European Commission proposes to ban certain uses of AI systems, it configures a number of risks that the European Commission (and other stakeholders) think are too high and subjects them to strong regulation, and it conceives other risks that are only subject to a minor regulatory stretch. The European Commission's reasoning is rather linear and can be summarised in the

following formula:

> **Mistrust because of AI risks → Identify, systematize and regulate AI risks → Trust as a result of enforcement of AI risks regulation**

In particular, the current proposal can be commented on from different perspectives. To start with, an explanation is lacking about the criterion or criteria laying behind the categorisation of risks. It is, for example, surprising that the risks of AI devices used for medical purposes are not contemplated in a specific way in the proposal. A second point has to do with the conformity assessment procedure. This procedure relies almost exclusively on the individual responsibility of the provider of the AI system. It is therefore a "private" conformity system assessment. To be sure, there are exceptions to the rule, but these exceptions are not as important as to trump the previous general rule. It is however unclear why high-risk AI systems are not subject to a third-party conformity assessment which would be in turn supervised by public administrations. Taking into account the specifics of AI devices, the treatment for high-risk AI systems should be completely different from the regulatory treatment that is currently given to regular products. A final, but not less important, point is that the difference between requisites and obligations is unclear. Some requisites at least imply obligations for the user and the provider of AI systems. This is not to say that a difference between requisites and obligations makes no sense; it only means that the demarking line between the two should be made clearer in the proposal.

The main argument of this paper can be summarised by saying that the current European Commission's proposal for the regulation of AI probably rests upon a misconception. Trust in AI systems will be impossible to achieve for the reasons that have been pointed out before—we do not understand AI systems' rationality and, what is possibly more important, the kind of rationality that AI systems will develop will not be similar to human-like rationality. In this context, it makes little sense to speak of trust in AI devices. Also, it makes no sense to try to regulate risks for trust as a way to solve this problem. In fact, a good deal of structural mistrust in AI systems might be beneficial for all.

CITED BIBLIOGRAPHY

Saleh Afroogh, *A Probabilistic Theory of Trust Concerning Artificial Intelligence: Can Intelligent Robots Trust Humans?*, AI & ETHICS, June 2, 2022, at 1, 1, 13–14.

Yann Algan, *Trust and Social Capital*, *in* FOR GOOD MEASURE: ADVANCING RESEARCH ON WELL-BEING METRICS BEYOND GDP 283, 286, 302 (Joseph E. Stiglitz, Jean-Paul Fitoussi, & Martine Durand eds. 2018).

Yann Algan & Pierre Cahuc, *Trust, Growth and Well-being: New Evidence and Policy Implications*, 2 HANDBOOK OF ECON. GROWTH (forthcoming June 2013).

Tony Bradley, *Facebook AI Creates its Own Language in Creepy Preview of Our Potential Future.*, FORBES (July 31, 2017, 11:20 AM), *https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/?sh=79504d9f292c.*

Hyesun Choung et al., *Trust in AI and its Role in the Acceptance of AI Technologies*, INT'L J. OF HUM.-COMPUT. INTERACTION (forthcoming March 2022).

JOHN ELSTER, STRONG FEELINGS: EMOTION, ADDICTION, AND HUMAN BEHAVIOR 190–91 (Francois Recanati ed. 1999).

See generally ANTONIO ESTELLA, LEGAL FOUNDATIONS OF EU ECONOMIC GOVERNANCE (2018).

See generally RUSSELL HARDIN, TRUST 46 (2006).

Steven Lockey et al., *A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities, and Future Directions*, 54 L. HAW. INT'L CONF. SYS. SCI. 5464–67 (2022).

ROBERT D. PUTNAM, BOWLING ALONE: THE COLLAPSE AND REVIVAL OF AMERICAN COMMUNITY 144–45 (2000).

ROBERT D. PUTNAM ET AL., MAKING DEMOCRACY WORK: CIVIC TRADITIONS IN MODERN ITALY 169–70 (1993).

PETER SINGER, ANIMAL LIBERATION 47, 49 (2d ed. 1995).

Alan F.T. Winfield & Marina Jirotka, *Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence*, PHIL. TRANS. R. SOC'Y A., Aug. 21, 2018, at 1, 1, 10 (2018).