Blockchain, Facebook and a Polygraph

Paulo Vieira[1], Caroline Stockman[2] and Paul Crocker[3]
[1]University of Beira Interior, C4- Could Computing Center, Covilhã, Portugal
[2]University of Winchester, UK
[3]University of Beira Interior, Covilhã, Portugal
paulo.vieira@ubi.pt
Caroline.Stockman@winchester.ac.uk
crocker@di.ubi.pt
DOI: 10.34190/EAIR.20.037

Abstract: Blockchain is a technology that makes use of a set cryptographic primitives that allow it to store information by consensus of the nodes without using a trusted central party. Blockchain mechanisms also make the stored information unalterable, and allow it to be verified in a quick way. This technology allows us to relate intangibles like truth, consensus and security. This paper presents the development of a polygraph application that can work in social networks using the users of the social network previously registered in the application as the nodes of a blockchain. Using the Nash equilibrium and techniques of Artificial intelligence a mechanism of consensus is created that allows the nodes to register information in the ledger. Truth and accuracy is one of today's greatest problems in social media platforms and as such we develop and present an architecture to work in facebook as the first application of our blockchain. The architecture is ethical by design, inspired by the moral formalism of Immanuel Kant as truth, and lying, were central concerns to his philosophy. The principles of his ethics can be implemented into blockchain technology so we establish a technology which upholds truth as a formal duty between people, by detecting falsehoods spread on social media. This ethical coding would in principle enable society to protect the rationality and dignity of its citizens. This is a small step towards the possibility of greater ethics en-coded in the practices of our digital world, whilst we do discuss and recognise the many technical and conceptual complexities – such as the nature of 'truth' and 'ethics' in itself, especially given Kant's hard-lined philosophy. However, Kant's universality of thinking aligns well with the logic of computing; therefore ethical formalism appears to be a suitable first step in the exploration of philosophically inspired truth judgements in blockchain technology.

Keywords: Blockchain, Facebook, Artificial Intelligence, Security, Ethics, Kant

1. Introduction

Blockchain technology allows stakeholders to make decisions that are valid for the entire platform. On this basis, it records its data in a ledger. These records become an inviolable history and the whole process is carried out without a third-party institution to validation. The process of consensus is a key concern, and the backbone of this technology. This project is a part of a project that we had been developing named Epistemological BlockChain (E-BC) (Vieira, Crocker, De Sousa 2019). Our solution to the consensus problem is an automatic mechanism using the Nash Equilibrium that emerged from Artificial Intelligence (AI) techniques involving Machine Learning and other AIs in the decision process. This necessitates the need to investigate a successful partnership of ethics and computing. We propose the fundamental idea of 'ethics-in-code'. If we can code AIs to 'think ethically', its functioning enhances trust, transparency, and discernment – all key aspects of AI debates today. We design a solution using all this for an emergent problem of our society, the value of the truth in social networks. We specifically consider Facebook as one of the most prominent social platforms, which has also been involved in ethically suspect cases of truth. Thus we design a polygraph to be implemented in Facebook in an app working in blockchain principles with consensus mechanisms involving AIs to evaluate information and classify them.

A central question would be where to source the ethical principles which would come to inspire the technical developments. A first source is agreement in public opinion. For example, Awad et al. (2020) describe their moral machine experiment: taking autonomous vehicles as their topic, they present participants with the 'trolley dilemma', a typical case scenario discussed in moral philosophy debates. They gathered 40 million decisions from millions of people in 233 countries and territories. They record global preferences: the preference to spare humans over animals, the preference to spare more people over fewer people, and the preference to spare the younger person over the older ones. In contrast, an committee of AI and policy experts in Germany drafted an

ethics code for autonomous vehicles (Luetge, 2017) – which, for example, prohibits discrimination based on age. So personal preference and expert opinion could conflict; but equally, there is a directional pull of the law on public opinion. Using the same topic of autonomous vehicles, Huang (2020) critically discusses 'law's moral halo', which is that people make different moral judgements on the basis of their knowledge of the legal implications. If the law disapproves, an individual is more likely to disapprove as well. Concluding their study, Awad et al. (2020) explicitly call for an interdisciplinary framework for the regulation of 'moral machines', which we introduce here. An alternative to either public opinion, expert agreement, or legal regime, is to search for guiding thought in the many moral philosophies we already have at hand.

Immanuel Kant provides the philosophical framework for this project. Certainly, there are critical considerations here which we highlight and keep under close scrutiny. But the principles of the categorical imperative, in their unwavering rigour, bring a sense of certainty to moral decision-making which would appear well-suited to transfer into code. There are many other moral philosophies available to continue this exercise, for example in consequentialist thought. While it would not be impossible for a machine to calculate consequences and determine the moral action on that outcome, that may be a matter of quantum computing - with this also ignoring the ever-looming danger of unintended consequences. A calculation of consequences is also not the basic intention of blockchain, the particular technology under consideration here. Blockchain aims to file and organise, and make available, information.

2. Kantian Ethics

2.1 Kant's Philosophy

Kant's moral philosophy is based on the virtue of reason, which he believes can set out to construct an ethical framework which applies to all. It is independent from local law, individual self-interest, personal feelings or circumstances – and to him, this is of the highest importance. If we do not have some kind of general, a priori framework, then we come to rely on localised judgements, which can certainly be good by accident, but this is not certain due to messiness of the practical human context (Kant, 1785/1993, p.22). He seeks a purer point of view, which can then be applied to empirical contexts, without becoming dependent on them. To Kant, something is good in itself, not because of its effects or outcomes (Kant, 1785/1993, p.7). True moral worth, in his words, follows from dutiful compliance with the categorical imperative. It is the non-negotiable compass for human action, and it applies a priori. (Kant, 1785/1993, p.29).

So, Kant's moral philosophy takes the action itself as the locus of morality. Moral decision making can be ensured by following the moral principles that steer the action in itself, regardless of the expected or actual consequences of the action. An action is moral or immoral when checking the drivers for that action. Drivers like 'incentive' and 'motive' are suspicious in Kantian ethics, as they imply personal interest. That was Kant's objection to utilitarianism, which postulates ethics as subject to individual pleasure and happiness. Instead of 'inclination', we must follow 'duty' (Jensen, 9134). This is deontological ethics, and dramatically different to consequentalism, for example. In that sense, Kantian ethics are not subject to individual contexts. The principles apply universally, much like blockchain, which operates separate from individual negotiation.

2.2 Principles of Action

Kant formulates a number of maxims to explain the normative decision-making steered by his moral philosophy. These are the universal maxims, or rules that should direct any intentional action. While they are distinct, they still form iterations of the same moral stance and are therefore closely related. These principles culminate in the idea of the 'kingdom of ends'. In relation to blockchain, it could be interpreted to mean a communal network based on the same universal principles. Each agent in the network (human or machine) is a fully rational decision-maker with equal status to all other agents in the community. It all 'hangs together' in normative moral harmony.

2.2.1 Universality

Firstly, Kant directs us to 'never act except in such a way that I can also will that my maxim should become a universal law' (Kant, 1785/1993:14). In other words, only do something if you would want everyone to do that all the time. ***The action can only take place if it would always occur, regardless of individual circumstances (Ax 1)***.

It implies the machine would perform an action, and always the same action, separate from individual human input. For example: Tay, an AI chatbot developed by Microsoft and released on Twitter in 2016, turned into an

offensive and hateful machine within 16 hours of its release. The algorithm learned from human interactions and mimicked these accordingly. Unfortunately, it had been immediately targeted by human users with overwhelming inflammatory content. The principle of universality supersedes the localised input. Arguably, one could still construct a 'LIEBOT' or Munchausen machine, on the basis of this principle, but Kant upheld a distinct emphasis on truth-telling over lies (Bendel, Schwegler & Richards, 2017). (There is the deeply problematic question on the definition of 'truth' at the heart of this issue, which is beyond the scope of this paper.)

2.2.2 Dignity

A second maxim of equivalent importance is that a person 'is not a thing and hence is not something to be used merely as a means', instead, people 'exist as ends in themselves' (Kant, 1785/1993:36). This principle dictates that we ***recognise everyone in our community as equal and worthy participants (Ax 2)***.

This means not simply enlisting them to suit my purpose, but ensuring they can make and act upon their own choices. For example, data protection laws have come into effect to regulate data harvesting for any kind of purpose, commercial purposes notably. We must protect the human capacity to self-steer behaviour on the basis of rationally negotiated decisions (linking dignity to the next maxim, autonomy). Technology cannot therefore morally turn a person into a data point for commercial exploitation beyond their knowledge or agentic rights. ***A software which treats people as a means to an end, is an immoral tool (Ax 3).*** To a broader interpretation, humanity cannot be a tool to serve a purpose; it is its own purpose. Technology must seek to protect that sense of dignity.

2.2.3 Autonomy

The supreme principle of morality, as Kant (1785/1993:44) describes it, emphasises the human ability to self-steer decision-making, or 'autonomy'. Here, human will is free to exercise the virtue of reason, bound by the maxims of universality and dignity.

Of course, as autonomous beings, it would seem intuitively possible that we author our own laws. We *can* choose to follow non-rational desires and inclinations (for example an action simply because 'we feel like it' or because it adds to selfish pursuits), but those conflict with the maxims above which formulate the categorical imperative. Kant's philosophy does not condone non-rational behaviour as autonomous, because it does not begin from the place of goodwill and respect which Kant emphasises as fundamentals of the categorical imperative.

We would not be free, agentic, autonomous beings if we simply followed impulses or pursued pleasures which disrespect others as ends in themselves. In view of blockchain, it would therefore be perfectly possible for a technology to be ethical by encoding universal laws, which are by their nature non-natural. Laws which have been authored by free, autonomous agents – human beings, who have universality and dignity in mind as guiding, but non-negotiable, principles. The principle therefore recognises the steering capacity of the AI developers.Similarly, all machines are coded with some kind of a priori framework, which limits 'autonomy' strictly and sheds a different light on the potential misconception of, for example, 'autonomous' vehicles. The vehicles do not act completely of their own accord; they are coded to calculate, consider, and execute without direct human input. Autonomy is not boundless, but constituted within a limiting framework. ***If a software adheres to the categorical imperative embedded in its code, its actions can be considered autonomous, but its duty is also to protect the autonomy of all others in the network (Ax 4).***

2.3 Critiques

A consequentialist critique is of course that we cannot judge actions in themselves as moral or immoral without considering the outcomes. For example, Togelius (2011) describes the difficulty in defining operational rules in a game design based on Kantian ethics. It reviews the interaction of human player actions with the game engine rules, and concludes it's tricky to find a good procedural balance.

There is a simplicity to Kant's theory which is attractive, but also problematic in practice. The rigid nature of the principles does not allow for discussion or negotiation. That may actually be helpful to technological decision making at this initial stage. Bringing ethical thinking into code is a novel way of programming, and structured, pure theory will be a necessity to make it through the first step. However, as we begin to understand the how-to of this process, and the capacity of technology grows to deal with messy human realities, a more flexible philosophy would be interesting to implement. One could say it will even be desirable, to avoid futures which are unfit to meet dynamic human ethics.

Immanuel Kant is also known for the questionable nature of certain writings, for example the use of his theory in defence of racism (White, 2013), or his condemnation of same-sex relationships (Schaff, 2001). This sheds a dark perspective on the integrity of his moral philosophy. However, White (2013) concludes Kant's *Observations on the Feeling of the Beautiful and Sublime* is a comedic response to Hume's theory, a parody, rather than a literal racialisation of the human 'species'. Similarly, Schaff (2001) discusses Kant's arguments on same-sex relations, and then reviews his moral theory to conclude that actually they are moral, but Kant's interpretation was skewed due to mediating factors of history and religion. Some argue the usefulness of a philosophy may not be deterministically linked to the person who theorised it. For example, Heidegger remains a substantive and significant philosopher, while his ties to Nazism are evident (Ellenberger, 2018). Jean-Jacques Rousseau, celebrated in the theory of childhood development as an enlightened thinker, is also convincingly claimed to have left his four children in orphanages (Mendham, 2014). So it would seem we would lose many of our leading philosophies if we dismiss their thought, because of the flaws of the person behind the thought.

However, it is too easy to excuse this as: 'philosophers were only human after all'. As we equip machines with moral thought, every angle of our own decision making should be carefully scrutinised. Racist soap dishes exist already (Scott, 2019). We must be vigilant not to also encode the invisible, pervasive technologies with such functionalities (which would arguably be even more dangerous because of its invisibility). Simplicity in this experiment was a requirement, to take these first steps in trying if this could even work at all, but closer scrutiny is needed to ensure we do the right thing for humanity in the long run. There could be extreme danger in guiding a machine to adopt unwavering ethical beliefs. Human judgement is valuable by its nature to be flexible and creative – human history has demonstrated the darker times of society, when judgement did acquire that radical hue. This exercise in coding will therefore bring about the far harder questions: which ethical philosophies become the steering principles of society, on a global scale? Could it be localised? What about individual thought and freedom of choice, or is this an international safeguarding issue which overrides personal preference? Should there be law-making around the transparency of ethics-in-code? What are the human rights in case of ethical conflict with the code? What are the basic, but global, ethical rules and responsibilities for coders and companies? All these questions (and more) echo the key concerns around AI today.

3. Ethic and Artificial Intelligence

Our working definition of AI follows the view outlined by the European Commission (2019), in that it is a computational system that carries out a task with an associate degree of perceived 'intelligence' beyond what could be called simple automated decision-making. While this is still a broad and challenging concept, it does imply a degree of agency is awarded to a digital technology in executing tasks independently from their human user. It may involve classification, prediction, recognition, … and any further action pertaining to data processing. It carries significant ethical implications. As the availability of AI technologies grow, a concern for ethics is growing exponentially too. A simple Scopus search shows 2012 peer-reviewed publications with the keywords 'Ethics' AND 'AI', since 1962 up to 2020 (may). Crucially, more than 675 of those are from 2015 onwards (Figure 1).

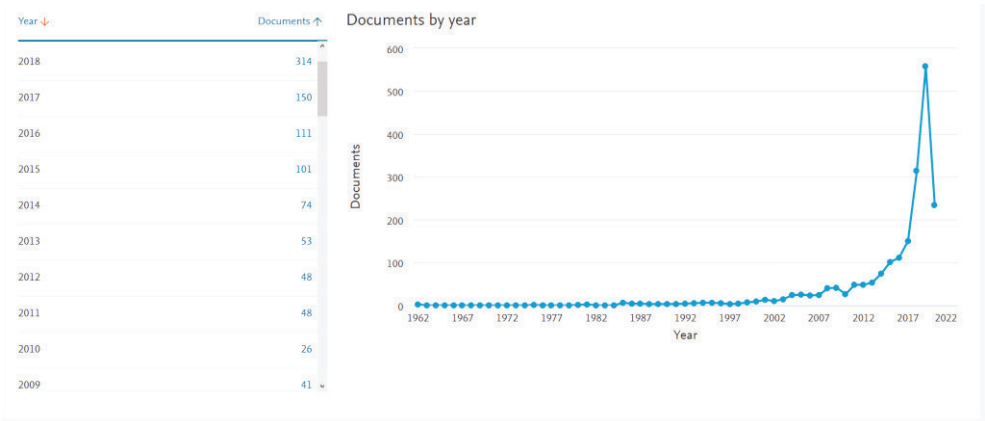


Figure 1: Results from scopus year by year (search 'Ethics' AND 'AI')

Analysing the abstracts of those publications closest to our purpose, we propose the following visualisation of the framework of thought surrounding ethics and AI (Figure 2). Our key concern is digital platforms regulation,

particularly Facebook on this occasion, to fight bias in digital information. Ethics is a driver to what should be done for AI development on this occasion. Ethics itself has been the subject of debate for centuries, however. What is an 'ethical' action? How do we come to morally justified decisions? If we're not sure ourselves, how can we tell our machines to think ethically?

So there is a choice to make, and a rather pressing one for that matter. The European Commission (2019:1) asks: 'how does an AI system achieve rationality?' As complex as this question is, Immanuel Kant's philosophy is prominently situated as one possible answer. His theory of moral action centres on rationality as the supreme virtue to ensure the ethics of decision-making.

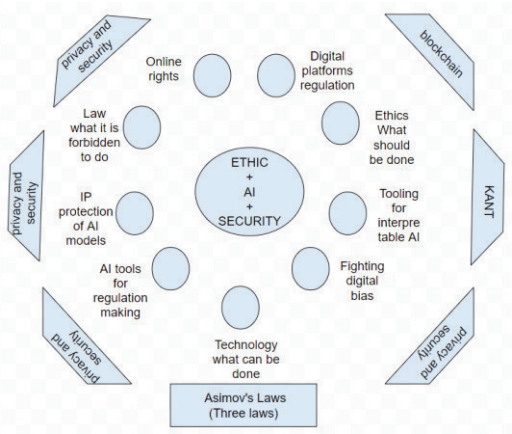


Figure 2: Ethic, Artificial Intelligence, Privacy and Security

4. A polygraph to social networks

4.1 Blockchain-Cloud Computing Architecture

An architecture Blockchain running using services of Cloud computing platform can be used when we design apps in platforms such as Facebook. Facebook is running on servers providing services in a logic layer of cloud computing. When we develop apps that can be used using their users it can be used in a BC logic. In this architecture BC logic the nodes are the users of the Facebook that join to be app users. The ledger is the history of the app, of the publications of the app, in Facebook. The publication in the ledger is done through a consensus mechanism that uses the Nash Equilibrium in a modal logic designed by us based on universality of the Kant Ethical thought. Facebook already have a platform based in knowledged with an ontology implemented in Graph whose architecture already is available to developers https://developers.facebook.com/docs/graph-api/

Table 1: Facebook representation knowledge graph to apps

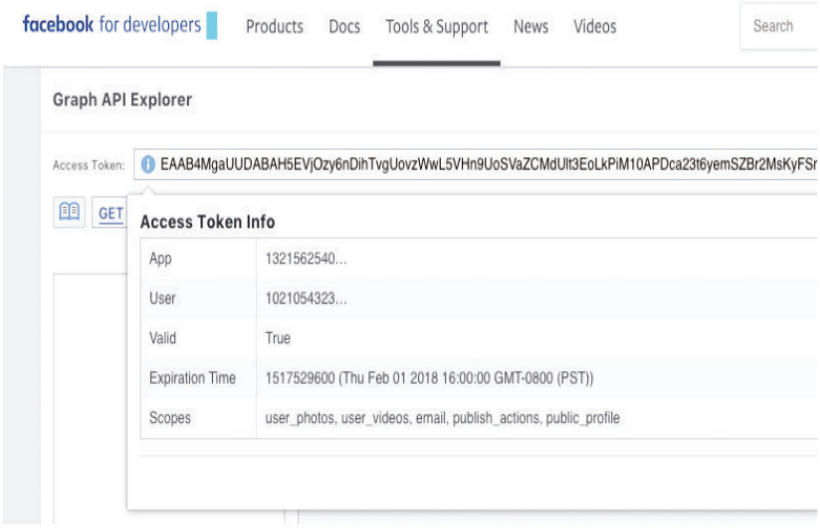
Graph - Nodes	
user	comment
photo	story
album	video
event	link
group	note

Graph - Edges	
Feed	Interests
Tagged	Likes
Posts	Photos
Picture	Stateuses
friends	Activities

Graph -field
User has: name, age, birthday, and so on
Page has: name, description, category, and so on
field Is the information about the User and Pages

How apps accessed to the platform information?
- HTPP based REST API
- The apps can be:
-> query data
-> post status ans storiesd
-> upload: pictures, videos, and more ...

Table 2: Graph API to developers (https://developers.facebook.com/docs/graph-api/reference)



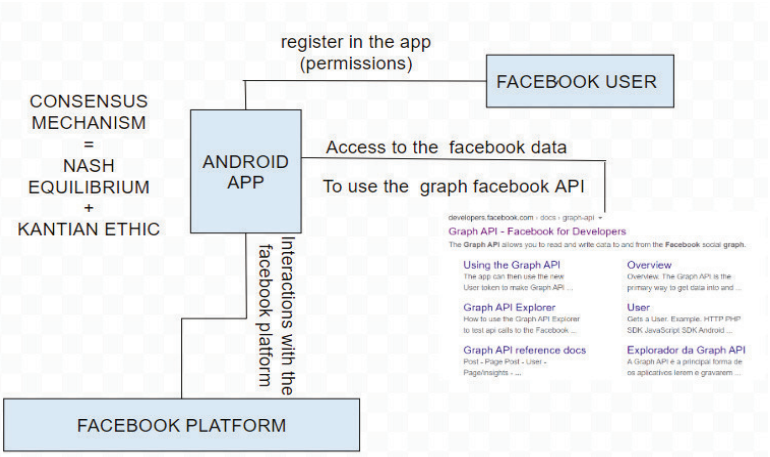
4.2 Nash Equilibrium and Kantian ethics

This is a framework of AI for governance and confidence in Social Networks Work in progress towards a polygraph block chain application that can be used for governance in social media and to increase trust and confidence in the material posted as well as to flag up fake news and other untruths. Our approach is grounded in the ideas of Kant and the immutable ledger offered by blockchain technology.

Our implementation of the E-BC as an app in Facebook will be done in Android technology[1], as the representation of information in Facebook is in graph objects, are graph data structures, we are in face of graph theory (Bondy 1976) as mathematical objects and data to programming. The mathematics have yet their presence throughout the logic of Kant thought will be implemented in a modal logic (Blackburn 2016) and by the Nash equilibrium that we implement is inheritance from game theory (Peters 2015). In really the Nash equilibrium and modal logic will be the mechanism of consensus of the E-BC app.

A Proof of Content, PoC, of the use of the Nash equilibrium in a BC logic of functionality already is done (Vieira, Crocker, De Sousa 2019). The PoC was performed through the implementation of code that was tested. This code simulated the functionality of the mechanism. Now we will add to this validation a working model logic principles of decision. The architecture design by Vieira, Crocker and Sousa in 2019 to the E-BC involved Machine Learning in the decision procedure, with this new addition the MLs decision (DIETZEN 1992) will be done in a model logic mechanism.

Table 3: The facebook polygraph architecture



[1] https://developer.android.com/training/basics/firstapp

The Kant Ethic are synthesized in the following sentences highlighted in the text above. We call them axioms[2]:

Axiom 1- The action can only take place if it would always occur, regardless of individual circumstances.
Axiom 2- Recognise everyone in our community as equal and worthy participants.
Axiom 3- A software which treats people as a means to an end, is an immoral tool
Axiom 4- If a software adheres to the categorical imperative embedded in its code, its actions can be considered autonomous, but its duty is also to protect the autonomy of all others in the network.

To these four laws we add more one that are a set of three that is left to us by Asimov (Anderson 2008).
Axiom 5- The three laws of Asimov

5. Conclusion and Future work

This paper is a continuation of the project to design a polygraph using tools inherited from blockchain technology. The first step describing the computation protocol is published in Viera, Crocker, & de Sousa (2019). Here, we have considered moral philosophy as the inspirational driver of concrete coding. Immanuel Kant provided the framework for this conceptual discussion. While there are certainly further critical considerations to make in view of the proposed ethics, it is perhaps encouraging to see the possibility of moral philosophy as a direct source of greater ethics in AI development. A next step of this project could consider hands-on development, in the form of a Facebook app which works in the way described in its initial foundation here. This would require writing the model logic for the Kant thought and implementing it on code. An exciting space, to be continued.

Acknowledgement

This work was supported by Operação Centro-01-0145-FEDER-000019 – C4 – Centro de Competências em Cloud Computing, co-financed by the Programa Operacional Regional do Centro (CENTRO 2020), through the Sistema de Apoio à Investigação Científica e Tecnológica – Programas Integrados de IC&DT.

References

Anderson, Susan Leigh. Asimov's "Three Laws Of Robotics" And Machine Metaethics. *Ai & Society*, 2008, 22.4: 477-493.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J.., Shariff, A. Bonnefon, J. And I. Rahwan (2018) "The Moral Machine Experiment", *Nature*, Vol 563, Pp 59–64.

Bendel, O., Schwegler, K. & B. Richards (2017) "Towards Kant Machines", *The Aaai 2017 Spring Symposium On Artificial Intelligence For The Social Good*, Pp 1-11.

Blackburn, Patrick; Van Benthem, Johan Fak; Wolter, Frank (Ed.). *Handbook Of Modal Logic*. Elsevier, 2006.

Bondy, John Adrian, Et Al. *Graph Theory With Applications*. London: Macmillan, 1976.

Dietzen, Scott; PFENNING, Frank. Higher-order and modal logic as a framework for explanation-based generalization. *Machine Learning*, 1992, 9.1: 23-55.

Ellenberger, W. (2018) *Het Tijdperk van de Tovernaars*, De Bezige Bij, Amsterdam.

European Commission (2019) *A Definition of AI: Main Capabilities and Disciplines*: *Definition developed for the purpose of the AI HLEG's deliverables,* https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Hern, A. (2016) "Microsoft scrambles to limit PR damage over abusive AI bot Tay", [online], *The Guardian*, https://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay.

Huang, B. (2019) "Law's Halo and the Moral Machine", *Columbia Law Review*, Vol 119, No 7, pp 1811-1828.

Jensen, O.C. (1934) "Kant's Ethical Formalism", *Philosophy*, Vol 9, No 34, pp 195-208.

Kant, I. (1785/1993) *Grounding for the Metaphysics of Morals*, Hackett Publishing Co, Indianapolis/Cambridge.

Luetge, C. (2017) "The German ethics code for automated and connected driving", *Philosophy & Technology*, Vol 30, No 4, pp 547–558.

Mendham, M. (2014) "Rousseau's Discarded Children: The Panoply of Excuses and the Question of Hypocrisy.", *History of European Ideas,* Vol 41, No 1, pp 131:152.

PETERS, Hans. *Game theory: A Multi-leveled approach*. Springer, 2015.

Schaff, K. (2001) "Kant, Political Liberalism, and the Ethics of Same-Sex Relationships", Journal of Social Philosophy, Vol 32, No 3, pp 446–462

Scott, K. (2019) "Racist Soapdishes and Rebellious (?) Children: Towards Human/AI Cooperation", European Conference on the Impact of Artificial Intelligence and Robotics - ECIAIR 2019, 31st of October 2019 – 1st of November 2019, pp 297-202

[2] The notion of axioms needs meta mathematical analysis. We do here a lazy use of the word not a mathematical use.