Stay Human. The quest for Responsibility in the Algorithmic Society[1]

Guido Gorgoni

Dipartimento di Scienze Politiche, Giuridiche e Studi Internazionali
Università degli Studi di Padova
guido.gorgoni@unipd.it

Abstract: recent developments of Artificial Intelligence based on machine learning techniques through Big Data raise multiple ethical and legal concerns, all of which ultimately do turn around the issues of responsibility, which is increasingly invoked not as a remedy but as a character which shall shape the whole development process of AI as well as its functioning. The characters of AI, taken in its technical and social role, challenge some established ideas related to human agency, namely responsibility. Recently two scholars like Jack Balkin (director of the Yale Information Society Project he founded on 1997) and Frank Pasquale (author of The Black Box Society: The Secret Algorithms That Control Money and Information, 2015) proposed "new laws of robotics for the Algorithmic Society" inspired to Isaac Asimov's ones, but targeting the human agents behind the development and the use of AI. On the other side, Responsible Research and Innovation model has been proposed as a model for the responsible development of AI. Whilst the reference to responsibility is appealing, nevertheless the inflation of its disparate usages may obscure the meaning associated with it. This article wants to contribute to the understanding of the issues behind the idea of preserving the human character of responsibility when confronted to the risks of its dissolution induced by the increasingly relevant roles played by AI in our societies.

Keywords: *Algorithmic Society, Artificial Intelligence, Responsibility, Responsible Research and Innovation.*

[1] This research has been conducted in the context of the research project "DOR 2019: Identity, Responsibility and Rights in the era of Big Data and Machine Learning" funded by the University of Padova.

1. New "laws of robotics" for the "Algorithmic Society"

Nourished by Big Data, Artificial Intelligence (AI) is gaining an increasing role in assisting or substituting human agents in complex tasks of decision-making. The societal issues posed by machine-learning are discussed since decades, but the contemporary development of automated decision-making[2] represents a major societal concern nowadays given the changed societal context, in particular as regards the ethical, legal economical and widely societal implications of AI[3]. In the era of Big Data and machine-learning algorithms these issues gain a different shape and size, given they pervasiveness in almost all the aspects of our daily life. Big Data and algorithms are two sides of the same medal, they develop and grow together the one feeding the other so that it seems plausible speaking of a shift from the so-called "Information Society" to the "Algorithmic Society":

> We are rapidly moving from the age of the Internet to the Algorithmic Society, and soon we will look back on the digital age as the precursor to the Algorithmic Society. What do I mean by the Algorithmic Society? I mean a society organized around social and economic decision-making by algorithms, robots, and Al agents, who not only make the decisions but also, in some cases, carry them out[4].

Whilst traditionally algorithms were programmed by defining "by hand" their binding decision-making rules and weights, contemporary more complex algorithms increasingly dispose of learning capacities[5] which may induce to compare their performances to those of human agents.

The surprising classification and prediction performances of learning algorithms then suggest the idea that the machine really learned and possibly understood something, while it actually just optimized a (huge) number of parameters searching into a given (rich) set of solutions[6].

Unlike traditional computer programs whose explicit and controllable logic could be unveiled, new advanced algorithms can develop autonomously,

---

[2] The first European legislative act dealing with the – nowadays common – idea of "automated decision-making" was the French *loi relative à l'informatique, aux fichiers et aux libertés du 6 janvier 1978* , in particular article 2 González Fuster 2014, 65. The issue then was taken into account within the European Law by the Directive Directive 95/46/EC (article 15) and is now contemplated by the Article 22(1) of the European General Data Protection Regulation (GDPR), which establishes the right not to be subject to decisions based solely on automated processing, which produces legal effects or other effects significantly affecting the subject.

[3] Informatics Europe & EUACM 2018.

[4] Balkin 2017, 1219.

[5] Mittelstadt et al. 2016, 3.

[6] Crafa 2019, 45.

similarly to the human process of learning[7], enriching the set of instructions initially coded, in a way that the evolution of the code not only is not predictable in advance, but also in such a manner that it may not be possible to retrace it retrospectively. The computational power of these programs increases dramatically both the scale and the complexity of analysis used for taking decisions, but at the same time this work is surrounded by an increasing veil of opacity, which produces uncertainty in reason of the potentially problematic impacts of the decisions taken by algorithms, since algorithms "may advise, if not decide, about how data should be interpreted and what actions should be taken as a result"[8].

Within the context synthetically dressed here, the "datafication of personal information"[9] nourished by Big Data, by reducing personal identity to calculable and therefore computable informations, exposes both individual subjects and society as a whole to new forms of vulnerability:

> These digital constructions of identity and traits affect people's opportunities to employment, credit, financial offers, and positions. They also shape people's vulnerabilities to increased surveillance, discrimination, manipulation, and exclusion [...] Companies and governments employ all of this information creatively in ever new contexts of judgment, yielding ever new insights, judgments, and predictions. In this way, people's lives are subject to a cascade of algorithmic judgments that fashion identity, opportunities, and vulnerabilities over time [...] The central problem we face today, therefore, is not intentional discrimination, but cumulative harm to identity and opportunities[10].

This delegation of decisional activities with legal impacts to algorithms raises the concern of the respect of fundamental legal principles such as Human Rights and the Rule of Law, in one side, and on the other contributes to making increasingly problematic the distinction between natural and artificial persons, in particular at the juridical level of the definition of legal personhood[11].

Facing this context, there are numerous appeals to preserve the responsibility of human agents, both on the side of the software programmers or creators, which may be hidden in the shadow of the machine operations, as well as on that of the single users which may be overridden by the results of the machine decisions. This brings at the forefront the pleas

[7] Informatics Europe & EUACM 2018, 8.
[8] Mittelstadt et al. 2016, 1.
[9] Mai 2016, 193.
[10] Balkin 2017, 1235–36.
[11] Fagundes 2001; Pietrzykowski 2017.

invoking an increased accountability of algorithms and, accordingly, the transparency of the decision-making process, all of which challenge the core of the responsibility idea, at the same time putting forward the question of preserving the authenticity of some fundamental human traits we link to this idea, such as identity, agency and responsibility[12].

As much as the societal dimensions of AI become increasingly relevant, claims to a responsible development of AI in order to grant accountability of the decision-making process and the transparency of the attribution of responsibility increase, aiming at making "readable" and understandable the whole decision-making process performed by algorithms, otherwise opaque. Along this lines of thought, inspired by the famous three laws of robotics proposed by Isaac Asimov, Jack M. Balkin[13] proposed three new "laws of Robotics in the Age of Big Data", extended to Al agents and (machine learning) algorithms, configuring new specific duties and obligations of robots towards both the users and the wider community. Asimov's three laws of robotics are:

> First Law: a robot may not injure a human being, or, through inaction, allow a human being to come to harm.
> Second Law: a robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
> Third Law: a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws[14].

Unlike Asimov's laws of robotics, which where conceived as constraints embedded in the programming code of the robots, as they were supposed to regulate their behaviour, and therefore were directly embedded in the "brain" of the robot (Asimov called it the "positronic brain"), Balkin states explicitly that the new laws of robotics he proposes are directed towards human operators, in particular those who program the algorithms or use the artificial agents, since the central problem is not that of regulating the way they operate, but rather – one step forward – that of "regulating the regulators", that is the humans conceiving and using those artifacts, since they are supposed to face not a strictly technical problem but wider societal (ethical, political, legal) concerns:

> In the Algorithmic Society, the central problem of regulation is not the algorithms, but the human beings who use them, and who allow themselves to be governed by them. Algorithmic governance is the

[12] Hildebrandt 2019.
[13] Balkin 2017.
[14] Balkin 2017, 1217, quoting Asimov, "Runaround", in "I, Robot."

> governance of humans by humans using a particular technology of analysis and decision-making[15].

Indeed, poses Balkin, it is misleading considering that the laws should be addressed to artificial agents, as this corresponds to the "homunculus fallacy"[16], that is the idea that AI operates fully autonomously so that it is possible ascribe good or bad intentions to it, just as if there was a little person inside the program making them do good or bad things. Instead, AI is designed and employed by humans, which are hidden in the shadow of AI discourse:

> These technologies mediate social relations between human beings and other human beings. Technology is embedded into-and often disguises-social relations.
>
> When algorithms discriminate or do bad things, therefore, we always need to ask how the algorithms are engaged in reproducing and giving effect to particular social relations between human beings. These are social relations that produce and reproduce justice and injustice, power and powerlessness, superior status and subordination.
> The robots, Al agents, and algorithms are the devices through which these social relations are produced, and through which particular forms of power are processed and transformed[17].

This fallacy goes together with the "substitution effect", which makes people treat artificial agents as if they were humans, the only difference being that they are at the same time better (in terms of power and speed) and more limited (they perform well a limited set of operations but lack all the features typical of human judgement). Another crucial feature of the substitution effect is that it feeds the "homunculus fallacy" by encouraging "the projection of life, agency, and intention onto programs and machines. This also encourages the projection of responsibility from the humans using the algorithms to the algorithms themselves". This way of considering artificial agents shadows the existing social power relations behind a veil of apparent neutrality, whilst these technologies "become part of social relations of power among individuals and groups"[18].

As a consequence of this reasoning, Balkin proposed "laws of robotics of an algorithmic society" are very different from Asimov's "laws of robotics", since they are codified not in the form of binary code, but in that of legal

[15] Balkin 2017, 1221.
[16] Balkin 2017, 1223 ff.
[17] Balkin 2017, 1223.
[18] Balkin 2017, 1225.

rules and principles, as they are primarily directed to regulate the social use of AI by humans, and only indirectly to regulate the modes of operation of the artifacts, so that "the laws we need are obligations of fair dealing, nonmanipulation, and nondomination between those who make and use the algorithms and those who are governed by them"[19], since these are laws regulating the relations between humans, even if mediated and shaped by the use of data and algorithms. Accordingly, the three new laws of robotics are[20]:

1) With respect to clients, customers, and end-users, algorithm users are *information fiduciaries*.

2) With respect to those who are not clients, customers, and end-users, algorithm users have *public duties*.

3) The central public duty of algorithm users is to *avoid externalizing the costs (harms) of their operations* (called "algorithmic nuisance" in analogy with the pollution of the environment), which implies obligations of transparency, interpretability, due process and accountability.

Without following in greater detail the arguments Balkin develops about the three laws, here is interesting focusing on the plea in favour of the preservation of the human traits of identity, and the subsequent appeal to human responsibility:

> A central concern is how identity – the association of persons with positive and negative associations and traits – is constructed and distributed in the Algorithmic Society [...] people's algorithmically constructed identities and reputations may spread widely and pervasively through society, increasing the power of algorithmic decision-making over their lives. As data becomes a common resource for decision-making, it constructs digital reputation, practical opportunity, and digital vulnerability[21].

Harms to reputation, discrimination, normalization, manipulation, the lack of transparency and accountability are the side effects of algorithmic decision-making, which represents its social costs. On this line of thought Frank Pasquale, acknowledging that "the cornerstone of Balkin's proposal is to create obligations of responsibility in systems that do not necessarily share the human experience of intent"[22], proposes to add a fourth law of Robotics explicitly directed at human agents and invoking a "responsibility-by-design" approach in AI development, aimed not at regulating robotics

[19] Balkin 2017, 1226.
[20] Balkin 2017, 1227.
[21] Balkin 2017, 1236.
[22] Pasquale 2017, 1254.

post hoc, but at influencing *ex ante* its development by making identifiable, and therefore responsibilising, the programmers who "can no longer hide behind a shield of disruptive experimentalism to deflect responsibility"[23].

Whilst acknowledging the importance of establishing a regime of responsibility based on the (side) effects of the algorithmic decision-making, Pasquale argues that it will not be desirable relying only on responsibility intended as a mechanism of compensation (as indeed the cost-benefit analysis may be lead in very different – and biased – ways), and that it is essential "maintain deontological patterns of justification in the technology world to complement the utilitarianism of cost-benefit analysis"[24]. In other words, the issue of responsibility cannot be framed only as a matter of (social) calculation, as ultimately there is at stake the quality and the nature of fundamental human traits "we will not always be able to offer precise valuations of the alarm or apprehension we feel at certain algorithmic transformations of human social relations"[25]. Pasquale relies on the opposition, highlighted by Mireille Hildebrandt, between two pairs, namely meaning and action, in one side, and information and behaviour in the other:

> The study and practice of [Modern] law have thus been focused on establishing the meaning of legal norms and their applicability to relevant human interactions, while establishing the meaning of human action in the light of the applicable legal norms. Data-driven agency builds on an entirely different grammar, its building blocks are information and behaviour, not meaning and action. We need to face the possibility that this will drain the life from the law, turning it into a handmaiden of governance (that fashionable term meaning anything to anybody), devouring the procedural kernel of the Rule of Law that enables people to stand up for their rights[26].

In order to preserve a world still informed by the idea of regulating human behaviour through meaning, as Balkin's the three new laws aim at doing, it is necessary for Pasquale ensuring that creators of robots or algorithmic agents are traceable and therefore identifiable. Accordingly, he proposes to complement Balkin's first three laws with a fourth law:

4) A robot must always indicate the identity of its creator, controller, or owner[27].

This principle works indeed as a meta-principle underpinning the first three laws and complementing what Asimov called the "zeroth" law stating

[23] Pasquale 2017, 1248.
[24] Pasquale 2017, 1250.
[25] Pasquale 2017, 1251.
[26] Hildebrandt 2016, 2.
[27] Pasquale 2017, 1253.

that robots must not harm humanity. Even if robots and algorithms do evolve away from the values initially programmed as a result of their capacity to interact with the environment, continues the Author, "the original creator should be obliged to build in certain constraints on the code's evolution to a) record influences and b) prevent bad outcomes"[28].

Indeed, together with the idea of personal identity and responsibility, it is the idea of the law to be challenged by the algorithmic processes of decision-making: "explainability matters because the process of reason-giving is intrinsic to juridical determinations – not simply one modular characteristic jettisoned as anachronistic once automated prediction is sufficiently advanced"[29]. So, a "responsibility-by-design" approach focusing on human responsibility must complement the already existing models of security-by-design and privacy-by-design. Implementing this idea may require, like in Asimov's laws, hard-coding some principles in the artificial artefacts, such as logs, and to develop accordingly licensing practices that explicitly take into account the case of problematic outcomes produced by the machine. Nevertheless unlike Asimov's laws, these hard-coded principles are not directed to robots but to humans, recalling Balkin's metaphor, they target "the Rabbi behind the Golem"[30] and not this latter.

2. What kind of responsibility for AI development?

This appeal to responsibility is not new within the technological domain, and indeed it has been repeatedly invoked since some decades under various formulas such as "technology assessment", "stakeholder engagement", "ethical, legal and social implications of research (ELSA)", "midstream modulation of science", and, more recently and more explicitly, the idea of "responsible innovation", formalized in the European context as "Responsible Research and Innovation" (RRI), an idea which is widely advocated even though there are various opinion on its precise implications and applications[31]. As an approach aiming at introducing societal responsibility already in the design of innovation, RRI is variously defined; among the different characterization offered in the literature we could rely on this one:

> Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability,

[28] Pasquale 2017, 1254.
[29] Pasquale 2017, 1252.
[30] Balkin 2017, 1222.
[31] Burget, Bardone, and Pedaste 2017.

> sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)[32].

Irrespective of the different theoretical perspectives adopted and operational models proposed of the idea, some common traits of RRI are shared across its different conceptualizations, in particular the need of involving societal stakeholders at an early stage of research and innovation, openness and transparency of the research and innovation process and commitment to be responsive to broad societal concerns, which have to be integrated into the research activities in a broader sense, including funding and the broad institutional environment supporting it.

Whilst in one side the reference to responsibility is appealing, nevertheless the inflation of its disparate usages may obscure the meaning associated with it. Many concepts are implied by this reflection on responsibility, such as agency, identity and responsibility, which define the constitutive traits of the modern moral-legal subject. Here we will explore more in depth the characters of the different responsibility models which remain implicit behind the invoked appeal to a responsibility preserving human traits. Indeed, whilst it is more or less clear that this appeal to responsibility targets human agents and "human" traits of responsibility linked to the idea of evaluating actions and giving reasons for them – "meaning and action" in Hildebrandt's terms –, the extent to this appeal to responsibility is less clear, especially form a legal point of view, and it needs to be developed in order to dissipate some residual doubts ad to deploy its full potentialities.

3. Beyond liability and accountability

The meaning of the word "responsibility" is almost saturated by the ideas of liability and accountability both in the moral and in the legal field; within this latter, the responsibility for technological innovation is dominated by the model of risk management. In order to grasp the essence of the current appeals to responsibility it is necessary dressing a preliminary panorama of the different meanings associated to the responsibility idea. The moral and legal understanding of responsibility is characterised by the semantic dominance of the idea of answering, both it in the continental tradition, namely that connected to the legal positivism of Hans Kelsen, as well as in the analytic legal tradition represented by Herbert Hart.

[32] Von Schomberg 2011, 9.

In Kelsen's *Pure Theory of Law* (1967)[33] the concept of responsibility is presented as purely formal construction. Responsibility is thought as a purely formal concept, defined by the legal concepts of imputation and sanction:

> Imputation, which expresses itself in the concept of responsibility, is therefore not the connection between a certain behavior and an individual who thus behaves. . . . Imputation, implied in the concept of responsibility, is the connection between a certain behavior, namely a delict, with a sanction[34].

Similarly, in his work dedicated to responsibility Hart affirms:

> The original meaning of the word 'answer', like that of the Greek 'ἀποκρίνεσθαι' and the Latin *respondere*, was not that of answering questions, but that of answering or rebutting accusations or charges, which, if established, carried liability to punishment or blame or other adverse treatment [...] There is, therefore, a very direct connection between the notion of answering in this sense and liability—responsibility, which I take to be the primary sense of responsibility [...] The other senses of responsibility are variously derived from this primary sense of liability—responsibility and are connected indirectly with the relevant sense of answer in that way[35].

The traditional legal idea of liability-responsibility is focused on the designation of a "responsible" subject, which depends on the imputation of the negative consequences established by the law to a legal subject, which may not be necessarily the actual agent nor a human being. The responsible subject in this sense is the one who is obligated to bear the consequences of an event after its occurrence, and is therefore a formal notion. Indeed Kelsen affirms explicitly that the idea of a "legal person" is a purely artificial notion, a metaphor expressing a conundrum of rights and obligations:

> The physical or juristic person who "has" obligations and rights as their holder, *is* these obligations and rights – a complex of legal obligations and rights whose totality is expressed figuratively in the concept of "person." "Person" is merely the personification of this totality[36].

This abstract nature of the responsibility idea and of the concept of a legal person operates a sort of virtualization of the legal subject which favours the idea of granting legal personhood also to AI agents. Whilst in one side this is not completely new for the legal system since personhood is granted to

[33] Kelsen's legal philosophy can be considered as the "standard" doctrine of continental European legal culture of the first half of the XXth century (and even beyond).
[34] Kelsen 2009, 81.
[35] Hart 1968, 265.
[36] Kelsen 2009, 173.

legal persons since a long time, nevertheless in the case of of the attribution of legal personality to AI it is necessary distinguishing different profiles of the responsibility idea. Whilst extending liability to AI does not seem – in principle – problematic, things become more complex when we consider accountability, for the reasons illustrated above, and are even more complex when confronted when considering responsibility for wider societal issues like in the model of RRI, where responsibility is orientated towards preventing future harms and towards orienting societal choices, more than at retrospectively by giving an account of what was done or bearing the consequences of a sanction. The spreading of different significations of the responsibility idea contributes to the ambiguities of the various appeals to "responsible innovation", also because responsibility is an eminently contested subject not only as regards its precise meaning but also as regards its normative content.

In order to make sense of this, we will proceed by distinguishing different declinations of the responsibility idea, and subsequently analysing the pertinence of the RRI model for structuring the claims behind the three, or better for (or even five) laws of robotics for the algorithmic society.

4. Models of responsibility

In fact, responsibility is "a syndrome of concepts"[37] which are variously interconnected between them, and which have developed along different understandings (or paradigms) of the responsibility idea. According to François Ewald[38] along the historical development of the legal forms of responsibility we can distinguish three theoretical models: fault, risk and precaution.

The model of fault is the archetypical form of the responsibility idea, as we have seen, and it corresponds to the obligation to answering in the liability sense (being subjected to negative consequences, be they legal or moral) in reason of the connection with the action. By its very nature this model of responsibility looks at the past as responsibility is based on the judgement on a past action .

The model of risk developed at the end of XIXth century with the advent of industrial revolution (in particular as a way to address the increasingly relevant work accidents); the idea of a sanction for a fault of the agent is replaced with that of a compensation of the victim for a damage. Hence responsibility is declined along the idea of risk, managed through the

[37] Vincent 2011.
[38] Ewald 2001.

mechanism of social insurance, this way disconnecting responsibility from liability making compensation independent from the ascertainment of a fault. Within this conceptual paradigm, responsibility is oriented towards the future since its function is that of anticipating damages by means of risk calculation and subsequent management techniques.

The model of precaution is linked to the development of the idea of responsibility towards the future within the ethical reflection[39], subsequently bridged within the legal field under the form of the precautionary principle[40]. The idea of precaution has been elaborated given the problems posed by the scientific and technological evolution and the limits of the two former paradigms to adequately cope with it, since they presuppose either the possibility to identify an author or the action (in the case of fault), or , in the case of risk, the availability of information to calculate risks and to quantify responsibilities accordingly. Indeed, contemporary science and technological developments are often characterized by epistemic uncertainty surrounding the consequences they generate, which renders the attribution of fault extremely difficult (and almost useless) and at the same tine jeopardizes risk calculation. The precautionary approach links responsibility to uncertainty, focusing on the anticipation of responsibility through its preventive exercise rather than on its ascription ex post facto, by imposing a duty of care even in absence of scientific evidence, which has been translated into the legal domain for the first time in the principle 15 of the Rio Declaration on Environment and Development:

> where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation[41].

Within the European Union legislation the precautionary principle has been clarified in a communication of the European Commission adopted in February 2000[42], detailing the concept as developed in the EU Union and establishing the guidelines for its application[43]. The precautionary principle defines an approach to risk management whereby if there is the possibility that a given policy or action might cause harm, and if there is still no

[39] Jonas 1984.
[40] Boisson de Chazournes 2009.
[41] Principle 15 of the Rio Declaration on Environment and Development (Rio de Janeiro, 3-14 June 1992). UN Doc. A/CONF.151/26, vol I, annex I, 1992.
[42] COM/2000/0001 final.
[43] Consecrated as a general principle of European law by the EU Court of Justice, despite being originated in the context of environmental regulation, the precautionary principle is now enshrined in article 191 of the Treaty on the Functioning of the European Union, under the title dedicated to the Environment.

scientific consensus on the issue, the policy or action in question should not be pursued. The jurisprudence of the EU Court of Justice extended its application also in other fields, consecrating it as a general principle of the European law. In particular, the precautionary principle operates where uncertainties undesirable harmful consequences of scientific innovation cannot be prevented with the general rules and standards of risk governance, so that the situation requires a case by case decision. Thus, the precautionary principle does not introduce new forms of liability nor new criteria of risk assessment, but aims at *responsibilising* the relevant actors in cases of scientific uncertainty or controversy about future harms. Specifically, RRI refers and develops precisely this latter sense of responsibility.

The pleas in favour of new laws of robotics seem to go as well in a similar direction, as they target particularly the process of the development of robots and AI more than its outcomes. The meaning of this plea in favour of the new laws of robotics can be clarified with reference to the different meaning of the responsibility idea according its orientation in time. In order to grasp the different meanings of the responsibility idea we can follow the idea proposed by Uberto Scarpelli[44] suggesting that the meaning of responsibility oscillates like a pendulum between two different poles, which we could characterise as a passive one, corresponding to the idea of being held responsible, and an active one, corresponding to the idea of assuming responsibility[45]. This distinction in its turn imposes to differentiate between a responsibility orientated to the past, as it is usually understood in legal terms, and a responsibility orientated towards the future (which is more frequent in ethical terms):

> In a temporal sense, responsibility looks in two directions. Ideas such as accountability, answerability and liability look backwards to conduct and events in the past. They form the core of what I shall call "historic responsibility". By contrast, the ideas of roles and tasks look to the future, and establish obligations and duties—"prospective responsibilities," as I shall call them. Accounts of legal responsibility tend to focus on historic responsibility at the expense of prospective responsibility[46].

Retrospective responsibility entails a judgment over a situation made *ex post facto*, so that characterises responsibility as a reaction. Prospective responsibility, on the opposite, expresses the idea of assuming responsibilities for a certain situation overcoming the simple compliance with existing

[44] Scarpelli 1981.
[45] Bovens 1998.
[46] Cane 2002, 31.

rules and duties[47]. The distinction between retrospective and prospective responsibility does not concern only the temporal dimension but also confers to each one of them has a distinctive character. In its prospective sense responsibility is understood as an attitude more than as an obligation, and becomes meaningful as long as it is voluntarily assumed by the subject, than attributed by the law.

This sense of the responsibility is captured by the idea of responsiveness, which, in contrast with liability or accountability, refers to attitudes and subsequent behaviour that extend over and above legal requirements, placing responsibility away from the semantics of responding to a charge, considered, as seen, as its primary sense from which all the other shall be derived.

Sticking to this fundamental difference is of paramount importance in order to be able to both understand and deploy the potentialities behind the idea of a "responsible" innovation.

5. Conclusions

AI developments defy the core characters of a human (inter)action, and linked regulation, based on shared meaning, in particular at the ethical and legal level of the definition of responsibility. Replacing human intelligence with AI may reduce or even destroy the symbolic field of a reciprocal interaction based on mutual recognition, by substituting it with formal and functional definitions, which may still work in practice but at the price of disconnecting it form the reality they aim at representing[48], a price which is worthy reflect upon. Recent discussions on AI development originating from different contexts have drawn the attention on similar issues, namely the loss of human role and responsibility in the decision-making process and the increasing role played by pervasive and opaque automated decision-making processes, notably advanced machine-learning algorithms.

The idea of RRI has been explicitly taken into account in connection with the idea of a responsible development of AI as a way to introduce the systematic – and not episodic – consideration of wider societal concerns into the design phases of AI[49], in particular as a way of embedding into the practice of research some essential features of the RRI idea such as anticipation, reflexiveness and participation[50], in order to structure a general

[47] Cane 2002, 48.
[48] Crafa 2019, 47.
[49] Brundage 2016; Stahl and Wright 2018.
[50] Owen et al. 2013.

approach to research and development in contrast to develop sector-specific – and therefore inescapably contextual – approaches, which may not offer a more general guidance for the practice of research.

Indeed it is of paramount importance figuring out the way AI researchers orient themselves with regards to long-term considerations, as a critical element of what it means to be a responsible innovator in the AI field involves also a self-reflection on the role one plays within the broader innovation ecosystem, as well as on the intended role of the research (and subsequent artefacts), as well as of their possible impacts, on society[51].

It is clear that RRI may be usefully invoked as a way to shift the focus from the responsibilities of AI agents to the responsibility of human agents over AI development, but – whatever be the underlying governance framework – the crucial aspect is opening AI development to a genuine interdisciplinary dialogue[52], which begins by acknowledging the nature of human relations hidden in the shadow of machine operations.

References

Balkin, Jack M. 2017. "2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data Lecture." *Ohio State Law Journal*, no. 5: 1217–42.

Boisson de Chazournes, Laurence. 2009. "New Technologies, the Precautionary Principle, and Public Participation." In *New Technologies and Human Rights*, edited by Thérèse Murphy, 161–94. Collected Courses of the Academy of European Law. Oxford, New York: Oxford University Press.

Bovens, M. 1998. *The Quest for Responsibility. Accountability and Citizenship in Complex Organisations*. Cambridge: Cambridge University Press.

Brundage, Miles. 2016. "Artificial Intelligence and Responsible Innovation." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 543–54. Synthese Library. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_32.

Burget, Mirjam, Emanuele Bardone, and Margus Pedaste. 2017. "Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review." *Science and Engineering Ethics* 23 (1): 1–19. https://doi.org/10.1007/s11948-016-9782-1.

[51] Brundage 2016.
[52] Crafa 2019.

Cane, P. 2002. *Responsibility in Law and Morality*. Portland, OR: Hart Publishing.

Crafa, Silvia. 2019. "Artificial Intelligence and Human Dialogue." *Journal of Ethics and Legal Technologies* 1 (1): 44–56.

Ewald, François. 2001. "Philosophie Politique Du Principe de Précaution." In *Le Principe de Précaution*, edited by François Ewald, Christian Grollier, and Nicolas de Sadeleer, 29–44. Paris: Presses Universitaires de France.

Fagundes, Dave. 2001. "Note, What We Talk about When We Talk about Persons: The Language of a Legal Fiction." *Harvard Law Review* 114 (6).

Hart, Herbert Lionel Adolphus. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press.

Hildebrandt, Mireille. 2016. "Law as Information in the Era of Data-Driven Agency." *The Modern Law Review* 79 (1): 1–30. https://doi.org/10.1111/1468-2230.12165.

———. 2019. "Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning." *Theoretical Inquiries in Law* 20 (1). http://www7.tau.ac.il/ojs/index.php/til/article/view/1622.

Informatics Europe & EUACM. 2018. "When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making." https://www.acm.org/binaries/content/assets/public-policy/ie-euacm-adm-report-2018.pdf.

Jonas, H. 1984. *The Imperative of Responsibility*. Chicago, IL: University of Chicago Press.

Kelsen, Hans. 2009. *Pure Theory of Law*. Translated by Max Knight. 5th ed. Clark, New jersey: The Lawbook Exchange.

Mai, Jens-Erik. 2016. "Big Data Privacy: The Datafication of Personal Information." *The Information Society* 32 (3): 192–99. https://doi.org/10/gfz43k.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2): 205395171667967. https://doi.org/10.1177/2053951716679679.

Owen, Richard, Jack Stilgoe, Phil Macnaghten, Mike Gorman, Erik Fisher, and Dave Guston. 2013. "A Framework for Responsible Innovation." In *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, 31:27–50.

Pasquale, Frank. 2017. "Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society." *Ohio State Law Journal* 78: 1243.

Pietrzykowski, Tomasz. 2017. "The Idea of Non-Personal Subjects of Law." In *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, edited by Visa A.J. Kurki and Tomasz Pietrzykowski, 49–67. Law and Philosophy Library. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-53462-6_4.

Scarpelli, Uberto. 1981. "Riflessioni Sulla Responsabilità Politica. Responsabilità, Libertà, Visione Dell'uomo." *Rivista Internazionale Di Filosofia Del Diritto* LVII (1): 27–79.

Stahl, Bernd Carsten, and David Wright. 2018. "Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation." *IEEE Security Privacy* 16 (3): 26–33. https://doi.org/10.1109/MSP.2018.2701164.

Vincent, Nicole A. 2011. "A Structured Taxonomy of Responsibility Concepts." In *Moral Responsibility*, edited by Nicole A. Vincent, Ibo van de Poel, and Jeroen van den Hoven, 15–35. Library of Ethics and Applied Philosophy 27. Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_2.

Von Schomberg, Rene, ed. 2011. *Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields: Presentations Made at a Workshop Hosted by the Scientific and Technological Assessment Unit of the European Parliament in November 2010*. Luxembourg: Publ. Office of the European Union.