European University Institute

ROBERT SCHUMAN CENTRE FOR ADVANCED STUDIES

EUI WORKING PAPERS

RSCAS 2020/63
Robert Schuman Centre for Advanced Studies

# Models of Law and Regulation for AI

Nicolas Petit and Jerome De Cooman

European University Institute

**Robert Schuman Centre for Advanced Studies**

# Models of Law and Regulation for AI

Nicolas Petit and Jerome De Cooman

EUI Working Paper **RSCAS** 2020/63

ISSN 1028-3625

Published in October 2020 by the European University Institute.
Badia Fiesolana, via dei Roccettini 9
I – 50014 San Domenico di Fiesole (FI)
Italy

Views expressed in this publication reflect the opinion of individual author(s) and not those of the European University Institute.

## Robert Schuman Centre for Advanced Studies

The Robert Schuman Centre for Advanced Studies, created in 1992 and currently directed by Professor Brigid Laffan, aims to develop inter-disciplinary and comparative research on the major issues facing the process of European integration, European societies and Europe's place in 21st century global politics.

The Centre is home to a large post-doctoral programme and hosts major research programmes, projects and data sets, in addition to a range of working groups and *ad hoc* initiatives. The research agenda is organised around a set of core themes and is continuously evolving, reflecting the changing agenda of European integration, the expanding membership of the European Union, developments in Europe's neighbourhood and the wider world.

For more information: http://eui.eu/rscas

The EUI and the RSCAS are not responsible for the opinion expressed by the author(s).

## Abstract

This paper discusses models of law and regulation of Artificial Intelligence (“AI”). The discussion focuses on four models: the black letter model, the emergent model, the ethical model, and the risk regulation model. All four models currently inform, individually or jointly, integrally or partially, consciously or unconsciously, law and regulatory reform towards AI. We describe each model’s strengths and weaknesses, discuss whether technological evolution deserves to be accompanied by existing or new laws, and propose a fifth model based on externalities with a moral twist.

## Keywords

Models of Law; Artificial Intelligence; Robotics; Regulation; Ethics; Risk Regulation; Externalities.

## Introduction

This paper discusses models of law and regulation of Artificial Intelligence (hereafter, "AI").[1] The goal is to provide the reader with a map of the law and regulation initiatives towards AI. Most law and regulation initiatives display, on their face, heterogeneity. Yet they often have common foundations. This paper surveys the four main model of law and regulation of AI that emerge (I), describes their strengths and weaknesses (II), discusses whether technological evolution should be addressed under existing or new laws (III), and puts forward a fifth model of law and regulation based on externalities with a moral twist (IV).

## 1. Survey of Law and Regulation Models for AI

In the literature, legal scholars and practitioners come to the question of law and regulation of AI through four mental models. That is the black letter law model (A), the emergent model (B), the ethical model (C) and the risk regulation model (D). We describe each of these models, and discuss specific applications to AI.

**Table 1: Description of Law and Regulation Model**

| | **Black Letter Law** | **Emergent** | **Ethical** | **Risk Regulation** |
|---|---|---|---|---|
| **Timing** | Reactive | Proactive | Proactive | Proactive |
| **Discussion** | Descriptive | Normative | Normative | Normative |
| **Approach** | Statutory and doctrinal interpretation = > what the law is | Normative => what the law should be | Teleological when deontological ethics<br>Ontological when consequentialism | Cost-benefit analysis, with possible precautionary principle |
| **Example** | Legal personhood and intellectual property | | Citizens' scoring and facial recognition | |
| **Issues** | Irrelevance | Redundance | Ethics lobbying<br>Ethical relativism<br>Ethics shopping | Knee-jerk regulation |

[1] AI, as a field, concerns itself with the construction of systems which are capable of rational behaviour in a situation. See Stuart J. Russel and Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd ed., Pearson 2010) 1-2 and 4-5. We acknowledge the fact that one must distinguish between artificial intelligence as a scientific discipline (AI) and artificial intelligence systems (AIS). While the former "includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)", the latter "are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal". See The European Commission's High-Level Expert Group on Artificial Intelligence, "A Definition of AI: Main Capabilities and Disciplines", Definition developed for the purpose of the AI HLEG's deliverables, April 8, 2019, available at <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341> accessed April 6, 2020, p. 8. This definition brings together artificial intelligence and (cognitive) robotics. Cognitive robotics refers to the "endowing of robots with more cognitive capabilities". See Stan Franklin, "History, Motivations and Core Themes" in Keith Frankish and William M. Ramsey (Ed.), *The Cambridge Handbook of Artificial Intelligence* (Cambridge University Press, 2014) 25. Furthermore, we discuss both AIs and robots under the same intellectual aegis, even though we acknowledge the differences between those two technological fields. We do this not only because we conjecture a degree of convergence between both technologies, but primarily because intelligent machines in soft or hard envelopes have the ability to "act upon the world". See Ryan Calo, "Robotics and the Lessons of Cyberlaw" (2015) 103(3) California Law Review, 513-564.

### *A. Black Letter Law Model*

In a black letter law model, the focus is on how existing laws apply to an AI system.[2] By black letter law, we mean of entire body of positive law, that is judicial and statutory law. In his seminal book *Theorie der juristischen Argumentation*[3], Robert Alexy explained that legal discourse tries to resolve the question of what is mandatory, allowed or prohibited by answering practical questions not in a general way but instead taking into account the restrictions driven by legal frameworks composed of binding norms.[4]

The black letter law model starts from the identification of the relevant law, to which it confront the matter of fact involving an AI system.[5] In practice, the matter of fact will often be an AI use case leading to a dispute.

In a black letter approach, the analysis is either conduct within a field of the law or across several fields of the law. In the first case, the disciplines most commonly looked at in the law and AI scholarship are product safety (including cyber security) and liability, consumer protection, intellectual property, labour law, privacy, civil liability, criminal liability, legal personhood, insurance and tax law. A particularly visible example of a disciplinary approach is the rights based approach. Under a rights based approach, a subset of legal and regulatory obligations pertaining to human rights, the rule of law and democracy are deemed so important that they become the main focus of inquiry in discussions over the law and regulation of AI systems.[6]

One disciplinary issue discussed under the black letter law model is whether AI-created inventions[7] or works of art can benefit from intellectual property ("IP") rights, and who is their owner.[8] Consider an

[2] This can be understood with a little green men metaphor. An alien from a distant planet sets foot on Earth. Most of Earth's clothing factories produce ready-to-wear suits for humans in calibrated sizes. Are human suits fitting, must they be stretched, adjusted, refitted? Or shall humans leave the alien naked?

[3] Initially published in 1978 under the name *Theorie der juristischen Argumentation: Die Theorie des rationale Diskurses al Theorie der Juristischen Begründung* and translated in English by Ruth Adler and Neil MacCormick in 1989. See Robert Alexy, *A Theory of Legal Argumentation: The Theory of Rational Discourses as Theory of Legal Justification* (Clarendon Press 1989).

[4] Matthias Klatt, "Robert Alexy's Philosophy of Law as a System" in Matthias Klatt (ed) *Institutionalized Reason: The Jurisprudence of Robert Alexy* (2012 Oxford University Press) 5.

[5] Robert Alexy, *Theorie der juristischen Agumentation: Die Theorie des rationale Diskurses als Theorie der juristischen Bedründung* (2nd ed, Suhrkamp 1991) 307-309. Alexy's theory however claims that some cases cannot be handled only on the basis of those norms due to the fact the lawmakers are sometimes unclear. Matthias Klatt, "Robert Alexy's Philosophy of Law as a System" in Matthias Klatt (ed) *Institutionalized Reason: The Jurisprudence of Robert Alexy* (2012 Oxford University Press) 6.

[6] Paul Nemitz, "Constitutional democracy and technology in the age of artificial intelligence" (2018) 376(2133) Philosophical Transaction of the Royal Society A: Mathematical, Physical and Engineering Sciences, Filoppo A. Raso et al, "Artificial Intelligence & Human Rights: Opportunities & risks" (September 25, 2018) Berkman Klein Center for Internet & Society at Harvard University, Karl M. Manheim and Lyric Kaplan, "Artificial Intelligence: Risks to Privacy and Democracy" (2019) 21 Yale Journal of Law and Technology 106, Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI" (January 15, 2020) Berkman Klein Center for Internet & Society at Harvard University, Mireille Hildebrandt, "The Artificial Intelligence of European Union Law" (2020) 21(1) 74-79.

[7] By AI-created invention, we mean an invention fully brought to existence by the AI, without human assistance.

[8] Christophe Leroux and al., "Suggestion for a green paper on legal issues in robotics" (*euRobotics*, December 31, 2012), <https://www.unipv-lawtech.eu/files/euRobotics-legal-issues-in-robotics-DRAFT_6j6ryjyp.pdf>: "Above mentioned legal IP systems are based on the fact that computers are inert tools, so that current intellectual property regimes usually only apply to humans or legal persons creations and not to creations coming from computers or inert tools. However, artificial technologies have advanced rapidly to the point that intelligent agents do not assist humans in the creation of works, but generate them autonomously. Thus, intelligent agents are capable of creativity". See also Ryan Abbott, "Hal the Inventor: Big Data and Its Use by Artificial Intelligence" in Cassidy R. Sugimoto et al (eds), *Big Data Is Not a Monolith* (MIT Press, 2016), David Levy, *Robots Unlimited: Life in a Virtual Age* (A K Peters, Ltd. 2006), 396-397 (hereafter Levy,

AI-created song. Under copyright law, courts insist that the work exhibits a “modicum of creativity”, reflects the “author’s intellectual creation” or constitutes the exercise of “non-mechanical, non-trivial skill and judgment”.[9] A debate exists today on whether the originality requirement prevents the allocation of copyrights to intelligent machines.[10] Similarly, in the area of patent law, an innovation is protected on condition that it involves an “inventive” step. As a rule, inventiveness means non-obviousness to a person skilled in the art. In layman’s term, a non-obvious discovery is one that is unexpected. But where should we set the benchmarks for “non-obviousness” and “skill in the art”, when one contemplates the introduction of AIs capable of “recursive self-improvement”?[11] What is not obvious to a man skilled in the art may be trivially evident for a super intelligent machine.

But the black letter law model also raises transversal issues that cut across various fields of the law. For example, can AIs be granted legal rights to litigate, contract or own property, including IP?[12] The parallels with discussions about the legal personhood of humans, corporations, international organisations and innate objects like trees are unmistakable.[13]

The black letter law approach is dominated by teleological questions. To solve fictional cases, courts and legislatures often consider the goals of the law.[14] For example, legal personhood was granted to corporations in order to promote economic exchange. A question that will therefore arise will be: should AIs be granted legal personhood to promote economic exchange, as was done for corporations? Similarly, a certain degree of legal personhood has been recognized to trees on grounds of sustainable development. In turn, it may be asked whether AI legal personhood is likely to contribute to the conservation of global resources.

### *B. Emergent Model*

The emergent model asks whether AIs raise new issues that require the creation of “a new branch of law”.[15] The assumption is that AI systems produce emergent phenomena.[16] Concretely, the emergent model asks whether AI systems’ unique economic[17], ethical[18] and scientific concerns require *sui generis*

*Robots Unlimited*) and Michael Gemignani, “Laying Down The Law To Robots” (1984) 21(5) San Diego Law Review, 1054.

[9] Elizabeth F. Judge and Daniel J. Gervais “Of silos and constellations: Comparing notions of originality in copyright law” (2009) 27(2) Cardozo Arts & Entertainment Law Journal, 375-408.

[10] A proxy is that a selfie taken by a monkey has been deemed unsusceptible of copyright protection because of lack of authorship. See, on this case, Joshua Jowitt, “Monkey See, Monkey Sue: Gewirth's Principle of Generic Consistency and Rights for Non-Human Agents” (2016) 19 Trinity College Law Review, 71-96.

[11] For use of this term, see Nick Bostrom, *Superintelligence, Paths, Dangers, Strategies* (Oxford University Press, 2014) (Hereafter Bostrom, *Superintelligence*).

[12] Samir Chopra and Laurence F. White, *A legal theory for autonomous artificial agents* (University of Michigan Press, 2011) 155 (hereafter Chopra and White, *Legal theory for autonomous artificial agents*) (conferring legal personhood necessitates a “decision to grant an entity a bundle of rights and concomitant obligations”).

[13] Christopher D. Stone, *Should Trees Have Standing? Law, Morality, and the Environment*, (3rd Ed Oxford University Press, 2010).

[14] See Chopra and White, *Legal theory for autonomous artificial agents*, 186 (“the decision to accord or refuse legal personality (both dependent and, in function of increasing competence, independent) would ultimately be a result-oriented one for courts and legislatures alike, and cannot rest solely on conceptual claims”).

[15] Levy, *Robots Unlimited*, 397. To use again our little green men metaphor (see *supra* n 2), the emergent approach is comparable to a tailor-made suit factory. The exercise consists in designing cloth that suits the alien.

[16] Ronald C. Arkin, *Behavior-based robotics* (MIT press, 1998), Ryan Calo, “Robots in American Law” (February 24, 2016), University of Washington School of Law Research Paper No. 2016-04.

[17] Ryan Calo, “Open robotics” (2011) 70(3) Maryland Law Review, 101-142 (hereafter Calo, “Open robotics”).

[18] Ethical arguments nurture demand for an all-out ban on research in relation to lethal automated weapons (“LAWs”). “Autonomous Weapons: An Open Letter from AI & Robotics Researchers”, July 28, 2015, <http://futureoflife.org/open-

legal prohibitions or exonerations of AI.[19] An emergent model is often at work behind books, articles and commentaries on "The law of driverless cars",[20] "The law of drones",[21] or "The law of robots".[22]

Often, the intellectual inquiry under the emergent model focuses on classes of AI applications. The *Stanford Artificial Intelligence and Life in 2030 Report* (the "Stanford Report") provides a good illustration.[23] The Stanford report purports to highlight how AI applications bring "specific changes affecting the everyday lives of the millions of people who inhabit them".[24] It focuses on eight applications domains where AI is deemed to have the greatest impact: "transportation, service robots, healthcare, education, low-resource communities, public safety and security, employment and workplace, home/service robots, and entertainment".[25] From this, the Stanford Report enumerates nine broad categories of legal and policy that AIs tend to raise: privacy, innovation policy, civil liability, criminal liability, agency, certification, labor, taxation and politics.

The Stanford Report displays commonalities, but also discrepancies with the black letter law model.[26] Some topics that were absent, irrelevant or subjacent in the black letter law model are prominent in the emergent model. This is the case of the legal arrangements governing certification (e.g., professional licensing requirements), taxation (e.g., how automated compliance reduces infringements to the law) and politics (e.g., voting and deliberation processes). And in the emergent model, the legal issues are framed as general topics that cut through several legal fields. Innovation policy is, for example, the umbrella framework under which liability issues, freedom of speech and patent law are discussed.

The emergent model is more focused on questions of *ex ante* legal design.[27] By this, we mean how to code an AI system to address a prospective legal issue.[28] This is distinct from the black letter law model, which focuses on legal arrangements for *ex post* frictional cases. One reason for this difference may be that technology-untrained lawyers are less comfortable discussing how to turn legal rules into computer code.

In addition, discussions under the emergent model are often more normative. Experts discuss "*Should* the law…" questions, while in the black letter approach they descriptively ask "*Does* the law…" questions. The emergent model thus gives a more explicit exposition to technological optimism or pessimism, which is often implicit in black letter law discussions.

letter-autonomous-weapons/>: "Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control". See also The European Commission's High Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI" (April 8, 2019) <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477> (hereafter: "AI HLEG Guidelines"). For more on the ethical approach, see *Infra* No I.C.

[19] In the computer science profession, there are concerns that "*overly rigid regulations might stifle innovation*", See The Economist, "You, Robot?" (The Economist, September 1, 2012). On this particular point, see *Infra* No III.A.

[20] See Alex Glassbrook, *The Law of Driverless Cars: An Introduction* (Law Brief Publishing, 2017).

[21] Michelle Bolos, "A highway in the sky: a look at land use issues that will arise with the integration of drone technology" [2015] 2 University of Illinois Journal of Law, Technology & Policy, 411-436; Tziporah Kasachkoff and John Kleinig, "Drones, Distance, and Death", in George J. Andreopoulos, Rosemary L. Barberet and Mahesh K. Nalla (ed), *The Rule of Law in an Era of Change* (Springer 2018), 15-45.

[22] Ryan Calo, A. Michael Froomkin, Ian Kerr, *Robot Law*, (Edward Elgar Publishing 2016).

[23] Artificial Intelligence and Life in 2030, One Hundred Year Study on Artificial Intelligence, Report of the 2015 Study Panel, September 2016.

[24] Ibid.

[25] Ibid.

[26] For instance, the Stanford Report stresses that AI is very relevant in relation to "regulation" without though discarding its relevance for other sources of law like common law, federal law, local statutes or ordinances.

[27] As explained above, this refers to the *ex ante* coding of legal rules in AIs and robots at the design stage.

[28] Olivier Boissier and al. "A roadmap towards ethical autonomous agents" (*EthicAa*, 2015) <https://ethicaa.greyc.fr/media/files/ethicaa.delivrable.3.pdf>.

The issue of whether AIs and robotic applications deserve legal rights helps understand what the emergent approach is concretely. Under this approach, one looks at technological outcomes, and applies a kind of Turing test to establish whether legal personhood can be granted. This approach was the one followed by Professor Lawrence Solum in a much-cited article when he considered the following thought experiment: "Could an artificial intelligence serve as a trustee?".[29] It is also the one followed when one asks whether an AI created song is approximates human art.[30]

Compared to the black letter law model, the emergent model is ontological. Since existing laws are often out of the picture, their goals are not considered. The intellectual inquiry focuses on understanding what the technology is. In the AI context, this is often done by reference to human intelligence.[31] The discussion focuses on a reflection on ourselves, and what makes us human.[32] And this is not neutral. Consider the idea of granting legal personhood to AIs. The emergent model may be biased in favor of anthropomorphic AI systems (robots like Asimo) or symbolic ones (softbots like Siri, Alexa or Cortana). Studies have shown that individuals treat computers like they behave with other human beings.[33] Evolutionary biology shows that people tend to treat as human what is like human. As Levy explained, "if our children see it as acceptable behaviour from their parents to scream and shout at a robot or to hit it, then […] our children might well come to accept that such behaviour is acceptable in the treatment of human beings."[34]

### *C. Ethical Model*

A third popular model focuses on ethics as the fundamental component of any law and regulation of AI systems. Ethics are the part of practical philosophy that deals with moral dilemmas. AI systems implicate mostly a field of ethics known as normative ethics.[35] The purpose of normative ethics is to create moral norms distinguishing the good and the bad. Applied ethics are also relevant to AI systems. Applied ethics analyze specific moral problems (e.g. abortion, euthanasia and now specific AI applications of like citizen scoring or facial recognition).

[29] Lawrence B. Solum "Legal Personhood for Artificial Intelligences" (1992) 70(4) North Carolina Law Review, 1231-1288.

[30] Nina I. Brown, "Artificial Authors: A Case for Copyright In Computer-Generated Works" (2019) 20(1) Columbia Science and Technology Law Review.

[31] Here, we fall in a complex philosophical discussion, as to whether our criteria of choice is Wittgenstein "family resemblance" or Aristoteles theory of definition (known as predicable doctrine), which focuses on genus and specific essence. See Michael R. Ayers "Locke versus Aristotle on natural kinds" (1981) 78(5) The Journal of Philosophy, 247-272.

[32] This issue has been thoroughly discussed by early philosophers since Aristoteles up until enlightenment. Those debates consist in a reflection of whether humans and animals are different by a matter of degree (as suggested by Darwin's theory of evolution, or are distinct in kind. Many properties have been underlined to denote the specificity of the human king: thought, language, instinct, self-consciousness, emotions, perfectibility, religion, vertical position, etc. For a good review (and rebuttal), see Philalethes. "The Distinction between Man and Animals" (1864) 2(6) The Anthropological Review, 153-163.

[33] See for example reference to computer-human interaction; Clifford Nass and Scott Brave, *Wired For Speech: How Voice Activates And Advances The Human-Computer Relationship* (MIT Press, 2005) 3-4. See also Byron Reeves and Clifford I. Nass, *The media equation: How People treat computers, television, and new media like real people and places* (Cambridge University Press 1996) and Mark Coeckelbergh, "Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations" (2011) 3(2) International Journal of Social Robotics, pp. 197-204.

[34] David Levy, "The Ethical treatment of artificially conscious robots" (2009) 1(3) International Journal of Social Robotics, pp. 209-216.

[35] There are three different branches of ethics: metaethics, applied ethics and normative ethics. See James Fieser, "Ethics", Internet Encyclopedia of Philosophy, <https://www.iep.utm.edu/ethics/>.

Normative ethics has three sub branches, ie virtue ethics, consequentialist ethics and deontological ethics. All three infuse debates over the law and regulation of AI systems.[36] Virtue ethics consider that happiness requires the practice of moral qualities in daily life. Aristotle singled out wisdom, justice, intelligence and moderation. Moreover, virtue ethics imply a "golden mean". Courage, for example, is the middle ground between cowardice and recklessness.[37] In an AI context, the requirement of transparency is an example of virtue ethics. And the middle ground reached by the requirement of explicability is a good illustration, because it requires some accountability, but not to the point of mandating exhaustive disclosure.

Deontological ethics consider that compliance with ethical duties determine whether an action is right or wrong, regardless of its consequences.[38] For example, assistance to homeless people (the intention) is rightful even if it implies a reduction of wealth for the benefactor (the consequence). Kant identified some "categorical imperatives".[39] A classical example given are the obligation to tell the truth or to treat humans with dignity.[40] The EU Guidelines on AI embrace a Kantian spirit when they state "*AI is not an end in itself, but rather a promising means to increase human flourishing*".[41] AI is a tool that should serve humanity, not the opposite. Deontological ethics sometimes lead to paradoxes. Always tell the truth is a categorical imperative. But what if a murderer rings at the door and asks an AI home assistant "where is your owner"?[42] The idea of ethical duty regardless of consequences is dangerous. The road to hell is paved with good intentions.

Consequentialism focuses on impacts. Rightfulness depends on a cost-benefit analysis.[43] Consequentialism is egoist when the costs and benefits are examined from the perspective of the agent that acts. It is altruist when the examination takes place from the perspective of society excluding the agent.[44] And it is utilitarist when the impact on society as a whole is considered.[45] A degree of consequentialism transpires from the European Guidelines on AI. One of their stated objectives is "to maximise the benefits of AI systems while at the same time preventing and minimizing their risks".[46]

The ethical model of law and regulation is technology neutral. Ethical recommendations adopted towards computational technologies like AI emulate solutions previously adopted in relation to biological technologies. Across the world, AI ethics tend to converge on the principles of beneficence, non-maleficence, autonomy, justice and explicability.[47] These principles come directly from bioethics,

[36] Jerome De Cooman, "Ethique et intelligence artificielle : l'exemple européen" [2020] 1 Revue de la Faculté de Droit de l'Université de Liège, 79-123.

[37] Aristotle, *Nicomachean Ethics*, translated by W.D. Ross (Kitchener, 1999).

[38] Markus Frischhut, *The Ethical Spirit of EU Law* (Springer 2019) 21(hereafter, Frischhut, *Ethical Spirit of EU Law*).

[39] Independently of our desires: "I ought never to act except in such a way that I could also will that my maxim should become a universal law". Immanuel Kant, *Groundwork of the Metaphysics of Morals* (1785), in Karl Ameriks and Desmond M. Clarke (eds) *Cambridge Texts in the History of Philosophy*, translated by Mary Gregor (Cambridge University Press, 1997), 15 (hereafter Kant, *Groundwork of the Metaphysics*). Kant was searching for rules we can universalise.

[40] Ibid., pp. 38-41

[41] AI HLEG Guidelines, 4.

[42] Helga Varden, "Kant and Lying to the Murderer at the Door… One More Time: Kant's Legal Philosophy and Lies to Murderers and Nazis" (2010) 41(4) Journal of Social Philosophy.

[43] Frischhut, *Ethical Spirit of EU Law*, 23.

[44] James Fieser "Consequentialist theories" Internet Encyclopedia of Philosophy, <https://www.iep.utm.edu/ethics/#SH2c>.

[45] Jeremy Bentham, *Introduction to the Principles of Morals and Legislation* (1789), reproduced in 2000 by Kitchener.

[46] AI HLEG Guidelines, 4.

[47] Luciano Floridi and Josh Cowls "A Unified Framework of Five Principles for AI in Society" (2019) 1(1) Harvard Data Science Review, 5 and Brent Mittelstadt, "Principles alone cannot guarantee ethical AI" (2019) 1 Nature Machine Intelligence, 501.

with the exception of explicability.[48] A debate exists today on whether it is right to draw inspiration from bioethics for AI systems. For example, while the interests of a patient and a doctor might be aligned, the same is not true for the developer of AI systems and its user.[49] Moreover, the regulatory environment of biotechnologies and AI differ widely, leading to distinct ethical requirements. For example, the regulatory framework in place in the health sector prevents hospitals from putting budgetary constraints before the interest of the patient.[50] By contrast, no such incentive constraint exists for AI.[51] Self-reliance on developers' willingness to respect ethical principles is key. Yet, research suggest that ethical statements have little or no impact on their daily practice.[52]

### *D. Risk Regulation Model and the Precautionary Principle*

A fourth model of law and regulation of AI systems is risk regulation. By risk regulation, we mean attempts to reduce the probability of occurrence *or* the levels of harms arising from events inherent in technology. In a White Paper on AI, the European Commission ("EC") takes a clear risk regulation approach when it calls for "a *regulatory framework* [that] *should concentrate on how to minimize the various risk* of *potential harm*".[53]

A risk regulation model of AI has several features. First, risk regulation proposes *ex ante* solutions. The goal is preventive, not corrective. Product design plays an important role, compared to insurance or liability. Red buttons, humans in the loop or sandboxing requirements in self-driving systems are a possible example.

Second, risk regulation mobilizes statistical evidence to evaluate risks.[54] For example, the German Data Ethics Commission proposed a pyramidal framework distinguishing five levels of risks, from the negligible (no special regulatory measure) to the existential one (complete or partial ban).[55] In that respect, risk regulation comes close to consequentialism and cost-benefit analysis: the higher the risk, the stronger the regulatory response.

That said, risk regulation's dependence on measurement encounters two limits. One, when calculation is impossible due to scientific uncertainty, precaution must prevail.[56] Absence of evidence does not mean evidence of absence. An event with uncertain probability but unsustainable consequences

[48] Ibid.

[49] Brent Mittelstadt, "Principles alone cannot guarantee ethical AI" (2019) 1 Nature Machine Intelligence, 501.

[50] Ibid., 502.

[51] Except maybe for particular activities, like privacy or data protection, which are protected in Europe through the General Data Protection Regulation.

[52] Andrew McNamara, Justin Smith and Emerson Murphy-Hill, "Does ACM's code of ethics change ethical decision making in software development?" (October 2018) Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 729-733. Brent Mittelstadt, "Principles alone cannot guarantee ethical AI" [2019] 1 Nature Machine Intelligence, 504 and Thilo Hagendorff, "The Ethics of AI Ethics – An Evaluation of Guidelines" (February 28, 2019) arXiv (Cornell University) <https://arxiv.org/abs/1903.03425>.

[53] European Commission, "White Paper on Artificial Intelligence: A European approach to excellence and trust", February 19, 2020, COM(2020) 65 final, p. 10.

[54] David Wright and John Copas, "Prediction Scores for Risk Assessment" in Robert Baldwin (ed), *Law and Uncertainty: Risks and Legal Processes* (Kluwer International Law 1997) pp. 21-38.

[55] Daten Ethik Kommission, "Opinion of the Data Ethics Commission (Executive Summary)" (October 2019) <https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2> p. 19 fig. 2.

[56] Roger Brownsword, *Rights, Regulation, and the Technological Revolution* (Oxford University Press 2008), 118-119.

is unacceptable.[57] This is where the precautionary principle gets in the game.[58] In AI contexts, a precautionary logic inspires calls for red lines, bans or moratoria on applications like lethal autonomous weapons, citizen scoring or facial recognition. Science is not irrelevant in the precautionary approach. Science helps establish the causal link between the event and its consequences. Besides, a precautionary approach is more than extreme consequentialism. The precautionary principle is a moral duty to ensure that everything possible is done to avoid catastrophic risk.

Two, cultural, political and psychological factors also influence risk regulation.[59] The Frankenstein complex, that is the Western fear that the creature (eg, an AI system) might one day outcompete and turn against its creator (humanity)[60] – a fear that Nick Bostrom calls the treacherous turn[61] – is strongly rooted in western culture, to the point that it appears in the European Report on civil law rules on robotics.[62]

## II. Four Fallacies of Law and Regulation for AI

The four models of law and regulation of AI exhibit dramatic shortcomings.

### *A. The Paradox of Irrelevant Law*

The paradox of irrelevant law concerns the black letter law model. Lawyers conjecture frictional rule-implementation cases with imperfect comprehension of the underlying technology. Because lawyers need case studies to apply their deductive discipline, they tend to rely on science-fiction to generate facts. Many scholarly works on AI and the law for example start with a discussion of Asimov's Three Laws of robotics. But because science-fiction is what it is[63], lawyers miss out on relevant technological evolution by focusing on fictional ones. The best example of this is driverless car. The dominant hypothesis in most science-fiction work prior to 2000 envisioned the period 2015-2025 with men and women driving flying hover cars, not a driverless one.[64] Had legal experts changed the law, we would today have a detailed, useless law of flying cars.

The black letter law model also leads to irrelevance due to blind spots. Reasoning from existing rules focuses our attention towards wrong directions. Our laws are abstract commands designed on the basis

[57] Joseph Norman and al., "Climate models and precautionary measures" (forthcoming), Issues in Science and Technology and Laurence Boisson de Chazournes, "New Technologies, the Precautionary Principle, and Public Participation" in Thérèse Murphy, *New Technologies and Human Rights* (Oxford University Press 2009), 168-169.

[58] Bridget M. Hutter, "The Attractions of Risk-based Regulation: accounting for the emergence of risk ideas in regulation" (March 2005), ESRC Centre for Analysis of Risk and Regulation, Discussed Paper No 33.

[59] Lukasz Gruszczynski, *Regulating Health and Environmental Risks under WTO Law: A Critical Analysis of the SPS Agreement* (Oxford University Press 2010) p.20.

[60] Sam N. Lehman-Wilzig, "Frankenstein unbound: Towards a legal definition of artificial intelligence" (1981) 13(6) Futures, 442-457.

[61] Bostrom, S*uperintelligence*, 144-145.

[62] "Whereas from Mary Shelley's Frankenstein's Monster to the classical myth of Pygmalion, through the story of Prague's Golem to the robot of Karel Čapek, who coined the word, people have fantasised about the possibility of building intelligent machines, more often than not androids with human features". Report of 27th January 2017 with recommendations to the Commission on Civil Law rules on robotics, (2015/2103(INL)), <http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html>.

[63] We acknowledge the fact that sometimes, science-fiction stories are anticipative. For instance, Jules Vernes wrote a story on a travel to the moon. It was science-fiction at the time, it is no more nowadays.

[64] With the notable exception of Isaac Asimov in Sally; Isaac Asimov, "Sally" (1953), Fantastic.

of specific representations of the state of the world, and its trajectories.[65] Too much reliance on the black letter approach undermines the necessary development of novel legal fields in a context of emergences.[66] One example illustrates the point. Assume that an AI ends-up dominating the world.[67] If this ominous prediction ever came true, should society introduce a "law of humans" which affords minority rights to humans and protects our species from intelligent machines? However, some of us cannot see this necessity today because all our laws embody the non-secular postulate that human conscience is special, and makes us superior to machines, animals and other entities. As a result of this cultural prior, our laws are essentially more about how humans treat non-humans (and in particular machines), and not about how non-humans (and in particular machines) treat humans.

### *B. The Problem of Redundant Law*

A fundamental aspect of emergent models of law and regulation of AI is to treat the technology as novel. As a result, emergent models of law and regulation often assume the existence of gaps in the law, or that AI systems operate in a lawless world.

Judge Easterbrook famously called this intellectual attitude the "*law of the horse*".[68] The expression derides the tendency to adopt new and *ad hoc* law when technologies emerge – in the XIXth century, the horse. Today, AI. The problem of the law of the horse is easy enough to see. Assume an AI assisted robot gardener causes a damage when mowing the lawn. Do we need to legislate specific rules on robot gardeners to address liability issues? The answer is a sure no. *Nove sed non nova* – not a new thing but in a new way.

In so far as AI is concerned, the law of the horse problem is essentially one of redundancy. AI is imitative.[69] Marvin Minsky wrote that AI is the "science of making machines do things that would require intelligence if done by men".[70]

If, on the one hand, law A governs human behavior while law B governs AI behaviour and, on the other hand, AI tends to imitate human conduct, then law A tends to copy law B. This is the problem of redundant law.

Most scholars that work under the emergent model tend to overlook the problem of redundant law, because they overestimate the capability or AI systems.[71] Human beings indeed have the tendency "to

[65] Take again the metaphor of the Alien (see *supra* n 2). His physiological condition may be different from ours. He may not need a suit at all. Instead, his physiological needs could be entirely different: for instance, he may be over sensitive to noise, and need soundproof helmets to cover its ears.

[66] Jack M. Balkin, "The Path of Robotics Law" (2015) 6(1) California Law Review Circuit, 45-60. ("The new technology disrupts the existing scene of regulation, leading various actors to scramble over how the technology will and should be used. As people scramble and contend with each other over the technology, they innovate—not only technologically, but also socially, economically, and legally—leading to new problems for law. Instead of saying that law is responding to essential features of new technology, it might be better to say that social struggles over the use of new technology are being inserted into existing features of law, disrupting expectations about how to categorize situations").

[67] Bostrom refers to this as the singleton scenario. See Bostrom, *Superintelligence*.

[68] Frank H. Easterbrook, "Cyberspace and the Law of the Horse" (1996) The University of Chicago Legal Forum, 207-216.

[69] McCarthy spoke about AI as the science of "programming computers to solve problems which require a high degree of intelligence in humans". John McCarthy, "Programs with common sense" (Stanford, 1959), <http://jmc.stanford.edu/articles/mcc59/mcc59.pdf>. He also made clear that "reaching human-level AI requires programs that deal with the common-sense informatic situation". John McCarthy, "Concepts of logical AI" (Stanford, 17 April 2000) <http://www-formal.stanford.edu/jmc/concepts-ai.pdf>.

[70] Marvin Minsky *Semantic Information Processing* (MIT Press 1968).

[71] Ted Striphas, "Algorithmic culture" (2015) 18(4-5) European Journal of Culture Studies, 395-412. Ian Bogost, "The Cathedral of Computation" (*The Atlantic*, 15 January 2015), <https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/>.

attribute programs far more intelligence than they actually possess (by any reasonably objective measure) as soon as the program communicates in English [or other natural language] phrases".[72] This bias feeds normative claims whereby new and *ad hoc* rules should be adopted for AI.

### *C. Ethics and the Failure of Good Intentions*

The ethical model is uncontroversial in its ambitions. Yet, it is rife with problems in its implications. The first set of problems is ethics lobbying (or ethics washing). Ethics lobbying arises when private or public organisations use ethical debates to prevent, delay or even replace legislation.[73] The idea is to rule out binding laws in favor of softer ethical rules better suited to technological innovations.[74] This concern has been voiced in the EU, whose Guidelines on AI have been criticized on the ground that they neither embody, nor call for, enforceable legal requirements.[75]

The second set of problems concerns ethical relativism. Put simply, there is no single ethics.[76] What one considers as good or bad is, by definition, personal. There is no objective way to rank Aristotle's virtues versus Kantian categorical imperatives.

The trolley problem illustrates ethical relativism (though some dispute its relevance)[77]. Consider a self-driving car, and place it in a situation in which there will be casualties. For example, the self-driving

[72] Massoud Yasdani, *Artificial intelligence: Principles and applications* (New York. Chapman and Hall, 1986) 326. See also Thomas J. Barth and Eddy Arnold, "Artificial intelligence and administrative discretion" (1999) 29(4) American Review of Public Administration, 348: "it is one thing to independently think through and analyze a question posed by one's superior; however, it is quite another to raise additional unasked questions or take the initiative and provide unsolicited advice where warranted. Even the most sophisticated AI system ultimately may be flawed because it lacks curiosity – that is, the urge to investigate issues or ask questions on its own".

[73] Thomas Metzinger, "Dialogue seminar on Artificial Intelligence: Ethical Concerns" (March 19, 2019) <http ://www.europarl.europa.eu/streaming/?event=20190319-1500-SPECIAL-SEMINAR1&start=2019-03-19T15:44:53Z&end=2019-03-19T15:56:00Z&language=en>, Nathalie Smuha, "The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence" [2019] 4 Computer Law Review International 101, Ben Wagner, "Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?" in Mireille Hildebrandt and Serge Gutwirth (eds.), *Being Profiling. Cogitas ergo sum* (Amsterdam University Press, 2018), Paul Nemitz, "Constitutional democracy and technology in the age of artificial intelligence" (2018) The Royal Society Publishing https://royalsocietypublishing.org/doi/full/10.1098/rsta.2018.0089, Luciano Floridi, "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical" (2019) 32 Philosophy & Technology, p. 188, Thilo Hagendorff, "The Ethics of AI Ethics – An Evaluation of Guidelines" (February 28, 2019) arXiv (Cornell University) https://arxiv.org/abs/1903.03425, Brent Mittelstadt, "Principles alone cannot guarantee ethical AI" (2019) 1 Nature Machine Intelligence, p. 501, Ryan Calo, "Artificial intelligence policy: a primer and a roadmap", (2017) 51 UC Davis Law Review, 399-436.

[74] Ben Wagner, "Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?" in Mireille Hildebrandt and Serge Gutwirth (eds.), *Being Profiling. Cogitas ergo sum* (Amsterdam University Press, 2018) 84 ("Ethical frameworks that provide a way to go beyond existing legal frameworks can also provide an opportunity to ignore them").

[75] Between the publication of the Draft (December 18, 2018) and the Report (April 8, 2019) on the European Ethics Guidelines, the document was open to a public consultation, in order to gather as many feedbacks as possible. All comments made during this open consultation are available at https://ec.europa.eu/futurium/en/system/files/ged/consultation_feedback_on_draft_ai_ethics_guidelines_4.pdf.

[76] For an illustration, see Floris de Witte, "Sex, Drugs & EU Law: The Recognition of Moral and Ethical Diversity in EU Law" (2013) 50(6) Common Market Law Review, 1545-1578.

[77] "Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, […] there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley?"; Judith Jarvis Thomson, "The Trolley Problem" (1985) The Yale Law Journal, Vol. 94, No 6, 1395-1415. Johansson and Nilsson argue that the trolley problem could in fact be solved by introducing self-driving cars. ("A key

car must choose between two options: 1. continuing its course, and killing a mother and her baby who crossed on red; 2. changing course, and killing an old woman on the sideway. How should the self-driving car, and the programmers in charge of optimizing algorithms, decide? Under the consequentialist approach, should it favor the younger or the older? Or should we randomize the decision of the self-driving car because it is what is closest to a human reaction.[78]

To answer these questions, the Massachusetts Institute of Technology (MIT) conducted a large-scale ethical study called the Moral Machine Experiment. The study sought to assess individuals' preferences in response to various death toll scenarios, e.g. when an autonomous vehicle should hit a wall to avoid a pedestrian who cross the road illegally.[79] The findings of the MIT study confirm the absence of universal ethics.[80] Three clusters of ethical values emerged, in the West[81], the East[82] and the South.[83] Individualistic cultures prefer sparing the many, while collectivist cultures tend to spare older members of the population[84]. And pedestrians who cross illegally have higher survival chances in countries which are "*poorer and suffer from weaker institutions, presumably because of their experience of lower rule compliance and weaker punishment of rule deviation.*"[85]

The third set of problems is ethics shopping. In a globalized world, companies can choose to locate their AI research operations in countries with weak ethical standards. In turn, companies can export weakly ethical AI systems to other jurisdictions, a practice known as ethics dumping.[86] As an importer of AI systems, the European Union was invited to rely on certification mechanisms to ensure that imported products have not been developed in countries with low ethical standards.[87]

enabler to disarm the trolley problem is the ability of the self-driving vehicle to estimate its own operational capability for handling surprising situations, and adjust its own tactical behavior accordingly.") Rolf Johansson and Jonas Nilsson, "Disarming the Trolley Problem – Why Self-driving Cars do not Need to Choose Whom to Kill" (September 2016) Workshop CARS 2016 – Critical Automotive applications: Robustness & Safety, Göteborg, Sweden.

[78] The outcome of such accident is more driven by instinct and human reflex than by a personal ethical debate. In fact, "human drivers who survive a crash may not have realized that they were in a dilemma situation". See Edmon Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Enrich, Azim Shariff, Jean-François Bonnefon & Lyad Rahwan "The Moral Machine experiment" (2018) 563 Nature, 59 (hereafter "The Moral Machine experiment").

[79] This experiment is available on <http://moralmachine.mit.edu/>.

[80] "This platform gathered 40 million decisions in ten languages from millions of people in 233 countries and territories"; Edmon Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Enrich, Azim Shariff, Jean-François Bonnefon and Lyad Rahwan, "The Moral Machine experiment" (2018) 563 Nature, 59 (hereafter "The Moral Machine experiment").

[81] The Western cluster contains "North America as well as many European countries of Protestant, Catholic, and Orthodox Christian cultural groups. The internal structure within this cluster also exhibits notable face validity, with a sub-cluster containing Scandinavian countries, and a sub-cluster containing Commonwealth countries"; "The Moral Machine experiment" 61.

[82] The Eastern cluster contains "many far eastern countries such as Japan and Taiwan that belong to the Confucianist cultural group, and Islamic countries such as Indonesia, Pakistan and Saudi Arabia"; "The Moral Machine experiment", 61.

[83] The Southern cluster contains "consists of the Latin American countries of Central and South America, in addition to some countries that are characterized in part by French influence (for example, metropolitan France, French overseas territories, and territories that were at some point under French leadership). Latin American countries are cleanly separated in their own sub-cluster within the Southern cluster"; "The Moral Machine experiment", 61.

[84] "The Moral Machine experiment", 62.

[85] Ibid.

[86] See Luciano Floridi, "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical" (2019) 32 Philosophy & Technology, p. 189, Charlotte Walker-Osborn, Callum Hayes "Ethics and AI a moral conundrum" (The British Computer Society, June 11, 2018), <https://www.bcs.org/content-hub/ethics-and-ai-a-moral-conundrum/> and Thomas Metzinger, "Dialogue seminar on Artificial Intelligence: Ethical Concerns" (March 19, 2019) <http ://www.europarl.europa.eu/streaming/?event=20190319-1500-SPECIAL-SEMINAR1&start=2019-03-19T15:44:53Z&end=2019-03-19T15:56:00Z&language=en>.

[87] This idea is especially proposed by the European Commission's independent experts on AI. See AI HLEG Guidelines, p. 23 and Luciano Floridi, "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical" (2019) 32 Philosophy & Technology, p. 190.

### *D. Knee-Jerk Regulation*

Knee-jerk regulation arises from an unwarranted application of the precautionary principle in response to realized risks and popular outcry.[88] The 2011 Fukushima nuclear disaster is a case in point. It combines a known "dread risk" (radioactivity), an intervening event (the disaster itself), and a reaction of stigmatization (essentially, by the public opinion).[89] In several Western countries, some have blamed nuclear technology and not Japan's exposition to seismic hazard.[90] As a result, some countries relatively unexposed to seismic and weather events have introduced nuclear exit policies, and turned to fossil fuel energies as an alternative.[91]

The potential for knee jerk regulation of AI systems is easy to foresee. Consider the case of a deficient AI airliner autopilot and assume that society displays a lower tolerance threshold for accidents caused by machines. In this context, it can be anticipated that society will respond to any crash with a prohibition of fully or partly AI-operated planes and roll back to require a significant degree of human operation. This, in spite of existing evidence that human operated flights may be significantly less secure than AI-assisted ones and that the source of the problem lies in the complex interaction between automated machines and humans.[92] Instead of prohibiting AI-assisted planes, regulation should seek to improve machine-human cooperation in ways that enhance safety.

The costs associated to knee jerk regulation also increase incentives on lawmakers to regulate anticipatively rare events with extreme impact. The problem, however, is that some rare events which have an extreme impact can only be explained and predicted after their first occurrence – Nassim Taleb speaks about "black swan" events.[93] Focusing on black swans to avoid them or, at least, be prepared, is a waste of time and resources.[94] Rather, "learning to learn" is more important.[95] After the Asian tsunami in 2004, the world had been awaken to dangers of undersea earthquakes. But it appears that not everyone drew the correct implication of undersea earthquakes and tsunamis for nuclear power plants until 2011. Perhaps is this due to the fact that, as Taleb note, we focus too much on specificities rather than generalities.[96]

[88] Also known as the "risk regulation reflex". Interestingly, knee jerk responses work also in the other direction. President Obama mocked the Republicans calls for deregulation as a knee jerk obsession: "Feel a cold coming on? Take two tax cuts, roll back some regulations and call us in the morning". See "Remarks by the President at the Democratic National Convention (September 7, 2012) <https://www.whitehouse.gov/the-press-office/2012/09/07/remarks-president-democratic-national-convention>.

[89] Emily Hammond, "Nuclear Power, Risk, and Retroactivity" (2015) 48(4) Vanderbilt Journal of Transnational Law, 1059-1082.

[90] William J. Kinsella, "Being "Post-Fukushima": Divergent Understandings of Sociotechnical Risk" (2015) Fukushima Global Communication Programme Working Paper Series.

[91] Speaking of the impact of the Three Mile Island accident in USA, Kranzberg note: "Yet the historical fact is that no one has been killed by commercial nuclear power accidents in this country. Contrast this with the 50,000 Americans killed each year by automobiles. But although antinuclear protestors picket nuclear power plants under construction, we never see any demonstrators bearing signs saying 'Ban the Buick'!"; Kranzberg, "Kranzberg's Laws", 552.

[92] David A. Mindell, *Our robots, ourselves: Robotics and the myths of autonomy* (Viking Adult, 2015).

[93] "What we call here a Black Swan (and capitalize it) is an event with the following three attributes. First, it is an outlier, as it lies outside the realm of regular expectations, because nothing in the past can convincingly point to its possibility. Second, it carries an extreme impact. Third, in spite of its outlier status, human nature makes us concoct explanations for its occurrence after the fact, making it explainable and predictable"; Nassim N. Taleb, *The Black Swan. The Impact of the highly improbable* (Random House 2007) xvii-xviii (hereafter Taleb, *Black Swan*).

[94] Ibid., xx-xxi.

[95] Ibid., xxi.

[96] Ibid.

## III. Law V Policy for AI

The normative question of whether technological evolution can be addressed by the justice system under existing laws, or whether it requires developing public policy, and in turn new laws, has long been discussed in the scholarly literature. In the early XXth century, Justice Holmes argued that the judicial process under the common law was apt to solve "*social questions*", and in particular, socio-scientific disputes, because what is really before the judge "is a conflict between two social desires, each of which seeks to extend its dominion over the case, and which cannot both have their way".[97] In contrast, and more recently, Spagnoletti has opposed that the legal system "*is inadequately prepared to cope with socio scientific disputes*",[98] in particular because adjudication exhibit interests through conflicts, and are thus ill suited to serve the public interest.[99] This issue can be conceptualized as whether one should have a dispute resolving or a policy implementing approach to the law and regulation of AI.[100]

### *A. Against Policy Implementing Approaches for AI?*

There are three arguments against policy implementing approaches of AI law and regulation. The first is that regulation stifles AI innovation. For example, data minimization requirements embodied in privacy regulation undermine the performance of AI systems. [101] It is common knowledge that AI systems rely one large datasets.[102]

[97] Oliver Wendell Holmes, "Law in Science and Science in Law" (1899) 12(7) Harvard Law Review, reprinted in Oliver Wendell Holmes, *The Collected Legal Papers* (Harcourt 1921) 210-43. See also, Elliott E. Donald "Holmes and evolution: Legal process as artificial intelligence" (1984) 13(1) The Journal of Legal Studies, 113-146 ("The image of the common law that Holmes presents in "Law in Science and Science in Law" is not of judges legislating rules. The architecture of the common law is not the product of conscious design choices by individual judges. It is rather the product of the logic of selection by a system, of an "invisible hand" like that of the market, or of natural selection in biology").

[98] Robert J. Spagnoletti "Using artificial intelligence to aid in the resolution of socioscientific disputes: A theory for the next generation" (1987) 2(1) Journal of Law and Technology, 101. Spagnoletti defines the legal system generally as a "method by which individual parties can present their interests and society can decide its own public policy".

[99] Ibid. Spagnolleti adds two other reasons. First, "[t]he present legal system, consisting primarily of nonscientific personnel, is unskilled in the relevant areas of expertise needed to understand the technology". Second, socio scientific disputes involve "quality-of-life decisions" and those are "difficult because they typically involve personal value judgments".

[100] Rafael La Porta, Florencio Lopez-de-Silanes and Andrei Shleifer, "The Economic Consequences of Legal Origins" 46(2) Journal of Economic Literature, 285-332.

[101] Ira S. Rubinstein, "Big Data: The End of Privacy or a New Beginning?" (2013) 3(2) International Data Privacy Law, 74-87; Omer Tene and Jules Polonetsky, "Privacy in the Age of Big Data: A Time for Big Decisions" (February 2012), 64 Stanford Law Review Online, 63-69.

[102] Randy Bean, "How Big Data Is Empowering AI and Machine Learning at Scale" (*MIT Sloan Management Review*, May 8, 2017), <https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/>. It is interesting to note that these concepts – AI and Big Data – are more than intertwined. Because big date gets bigger every day, AI is use to capture and structure big data. See Daniel E. O'Leary, "Artificial Intelligence and Big Data" (2013) 28(2) IEEE Intelligent Systems, 96-99. Some authorities consider the possibility to regulate large sets of data. For instance, Article 5(c) of the General Data Protection Regulation stipulates that "personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')". Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, L 119/1. See also Robert D. Atkinson, "IP Protection in the Data Economy: Getting the Balance Right on 13 Critical Issues" (January 2019), Information Technology & Innovation Foundation (ITIF), <http://www2.itif.org/2019-ip-protection-data-economy.pdf> "regulators began preventing companies from acquiring large amounts of data, this would delay or prevent many important technological advancements. For example, Tesla's self-driving technology, IBM Watson's ability to diagnose medical illness, and the Weather Company's weather predictions would all be impossible without massive amounts of data".

Such concerns have led Ryan Calo to support a specific immunity regime for AI and robotics manufacturers, close to the immunities enjoyed by firearms producers and website operators.[103] This means that AI systems should be safe, as designers remains liable under classic product safety laws, but are immune from lawsuits for improper uses of their products. Such a selective immunity constitutes a trade-off between safety and incentives.

The second argument against policy is that interest groups capture regulation.[104] In the context of AI, one area with strong rent seeking potential by private interest groups is car insurance. In many countries, the law imposes insurance duties on driver and/or user. With self-driving cars, the case for driver and/or user compulsory insurance is less compelling. There is less driver control, fewer accidents and lower damages at society level.[105] Of course, trees and snow still fall, causing casualties on the road. However, as autonomy progresses, allocating liability on driver and/or user seems less justified, and a transfer to driverless cars manufacturers is a more plausible option. Moreover, the insolvency concern that underpins the compulsory nature of insurance seems less problematic with car manufacturers.[106] The problem, of course, is that insurance companies have much to lose if compulsory driver and/or user insurance is abandoned. Their relative bargaining power against a handful of manufacturing companies is much lower than in relation to myriad individual drivers and/or users.[107] Last, car manufacturers exposed to hold-up conduct by insurance companies, may have incentives to vertically integrate into insurance services, rendering the insurance industry irrelevant in the long term. This situation incentivizes insurance companies to lobby in favour of an extension of compulsory driver and/or user insurance for self-driving cars.[108]

But public interest groups too can capture regulation. Public choice theory hints that government officials might favor technology that maximizes their own returns. Moses and Chan explain that instead of using the most useful AI technology, governments discharging law enforcement functions might focus on the most lucrative ones like technology that optimize fining – or those that contribute with its enforcement activities.[109]

The last argument against public policy is that regulation cannot keep pace with technological progress. When adopted, it may already be obsolete.[110] Gemignani however notes that both law and

[103] Calo, "Open robotics" ("Congress should shield manufacturers and distributors of open robotic platforms from suit for what consumers do with their personal robots, just as it immunizes gun manufacturers from suit for what some people do with guns and websites operators for what users upload and post").

[104] Anne O. Krueger, "The Political Economy of the Rent-Seeking Society" (1974) 64(3) The American Economic Review, 291-303.

[105] More so if the law incrementally reduces individuals' freedom to drive. See Dan McCrum, "Insurers will destroy themselves to nudge us into robot utopia" (*Financial Times*, March 4, 2014), <https://ftalphaville.ft.com/2014/03/04/1787962/insurers-will-destroy-themselves-to-nudge-us-into-robot-utopia/>.

[106] Michael G. Faure, "Economic criteria for compulsory insurance" (2006) 31(1) The Geneva Papers on Risk and Insurance Issues and Practice, 149-168.

[107] See "EU considers new insurance laws for driverless cars" (*Euractiv*, 2016) <https://www.euractiv.com/section/digital/news/eu-considers-new-insurance-laws-for-driverless-cars/>.

[108] By the same token, car manufacturers may lobby government so as to be insulated from all liability, and deflect it towards software developers.

[109] Lyria Bennett Moses and Janet Chan, "Using Big Data for Legal and Law Enforcement Decisions: Testing the New Tools" (2014) 37(2) University of New South Wales Law Journal, 659: "Chan's research found that the most successful use of information technology for proactive policing was in support of traditional law enforcement: the use of mobile data systems in police cars to check for outstanding traffic offence warrants. The enthusiastic adoption of this technology is easily explained by its effectiveness, as evidenced by 'an exponential increase in the collection of fines as well as the imprisonment of fine defaulters".

[110] Consider a proposed Federal Aviation Administration ("FAA") regulation that insists that unmanned aircraft system ("UAS") must be limited "to daylight-only operations, confined areas of operation, and visual-line-of-sight operations". Now compare this with surveys reporting that pilots spend just seven minutes manually operating their planes in a typical

public policy share this problem: "*the law generally reacts to issues only after they have become the center of a real controversy. Courts generally, and some courts exclusively, address a question of law only after an actual dispute involving that question has been brought before them. Legislation is also more often reactive than proactive. Yet in a society that seems to lurch from crisis to crisis, it is unclear whether such a strategy can avert eventual disaster. The danger of this reactive approach to technological advance becomes clear when dealing with robots and computers*".[111]

Yet, public policy raises specific regulatory "pacing" questions because lawmakers must decide when to intervene. This is not the case for courts. Gregory Mandell calls this issue a "quandary": whilst a lawmaker would like to discourage research in harmful technologies and incentivize research in beneficial ones, the risks and opportunities created by emerging technologies cannot be "suitably understood until the technology further develops".[112] The regulatory process must therefore keep a degree of "connection", and wait for technology to develop so as to endow the social planner with enough knowledge.[113] However, as the lawmakers acquire the necessary knowledge, the technology entrenches and it may be too late to act. This is known as the Collingridge paradox.[114]

This risk is discussed by Nick Bostrom as the "treacherous turn". This notion refers to the pivot point which is reached when a recursive self-improving AI becomes sufficiently strong to strike humans without warning or provocation.[115] In a matter of minutes, a malignant AI may consider that humans are threats to the achievement of its final values and turn against them avoiding the controls systems set by engineers. Bostrom uses the example of an AI designed to optimize production in a paperclip factory. Following a treacherous turn, the AI would proceed by first "converting the Earth and then increasingly large chunks of the observable universe into paperclips".[116] Elon Musk held a similar speech when he tried to convince that even a seemingly harmless AI system could have disastrous consequences.[117]

### *B. Against Existing Dispute Resolution for AI?*

The drawbacks of developing public policy for AI should not lead to the conclusion that reliance on courts is sufficient. Public policy is a necessity because the alternative – case-by-case dispute resolution

flight. See John Markoff, "Planes Without Pilots" (*The New York Times*, April 6, 2015) <http://www.nytimes.com/2015/04/07/science/planes-without-pilots.html?_r=2>. No wonder why companies like Amazon, Intel and Google have railed against emerging drones delivery regulation, which they consider outdated. See Ben Popper, "What's really standing in the way of drone delivery?" (*The Verge*, January 16, 2016) <https://www.theverge.com/2016/1/16/10777144/delivery-drones-regulations-safety-faa-autonomous-flight>.

[111] Gemignani, "Laying Down the Robot", 1046.

[112] Gregory N. Mandel, "Regulating emerging technologies" (2009) 1(1) Law, Innovation and Technology, 75-92.

[113] Roger Brownsword and Morag Goodwin, *Law and the Technologies of the Twenty-First Century. Texts and Materials* (Cambridge University Press, 2012).

[114] David Collingridge, *The Social Control of Technology* (Francis Pinter Ltd., 1980). Biotechnologies in the large sense are an example. Whilst they promise many wonders in public health, nutrition and environmental protection, the creation of new eco-systems could precipitate the demise of animal or human species, create new diseases, etc.

[115] Bostrom, S*uperintelligence*, 144-145.

[116] Ibid., 150.

[117] At a Vanity Fair's Conference, Elon Musk said: "Let's say you create a self-improving A.I. to pick strawberries, and it gets better and better at picking strawberries and picks more and more and it is self-improving, so all it really wants to do is pick strawberries. So then it would have all the world be strawberry fields. Strawberry fields forever"; Maureen Dowd, "Elon Musk's Billion-Dollar Crusade to Stop the AI Apocalypse" (*Vanity Fair*, March 26, 2017) <https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x>.

– is worse.[118] Prospects of litigation chill research and innovation incentives too[119]. For example, Ryan Calo has warned of the risk of crippling legal liability regimes in the field of open robotics.[120] The same applies to AI. Uncertain liability rules act as disincentives to investment, and channel the flow of capital towards narrow functionality where producers can better manage risk, leaving general AI robotics underdeveloped. And the uncertain application of existing legal institutions at early phases of technological development does not allow the formulation of safe appropriability propositions required to attract venture capital[121].

Besides, others insist on the potential of regulation to enable innovation. The Porter hypothesis states that strict environmental, health and safety standards prompt firms to improve their productivity, and finds that "properly designed [regulatory] standards can trigger innovation that may partially or more than fully offset the costs of complying with them".[122] In Porter's view, "tough standards trigger innovation and upgrading", and prompt firms to re-engineer. In addition, strict regulatory standards can promote market competition, by inducing firms to race for first movers' advantages.[123] In the AI field, Smuha mentions adoption of a fast-track migration policy for workers with an AI-related background.[124]

## IV. A Fifth Model? Externalities with a Moral Twist

Section I surveyed existing models of law and regulation for AI. This section describes a novel model. We propose to index the law and regulatory response upon the nature of the externality – positive or negative – created by an AI system, and to distinguish between discrete, systemic and existential externalities. The model brings together all existing models of law and regulation for AI in a consistent framework. Relating to section III, deviations from existing law towards the creation of AI specific law and regulation by public policy should be indexed on the type of externality generated by the technology. We introduce some key concepts first (A). We then discuss concrete applications (B).

[118] Epstein for instance poses the necessity of regulation: **"**At bottom, the proper inquiry never poses the stark choice of regulation versus no regulation**".** See Richard A. Epstein, "Can Technological Innovation Survive Government Regulation" (2013) 36(1) Harvard Journal of Law & Public Policy, 88.

[119] Product liability litigation in relation to deficient medical devices is an often-heard worry.

[120] Calo, "Open robotics".

[121] See Calo, "Open robotics": "legal uncertainty could discourage the flow of capital into robotics or otherwise narrow robot functionality".

[122] Michael E. Porter and Claas Van der Linde "Toward a new conception of the environment-competitiveness relationship" (1995) The journal of economic perspectives, Vol. 9, No 4, 97-118.

[123] Nicholas A. Ashford and Ralph P. Hall. "The importance of regulation-induced innovation for sustainable development" (2011) Sustainability, Vol. 3, No. 1, 270-292. Pelkmans and Renda document empirical examples of enabling regulation. One of them is the regulation of end-of-life vehicles. Under the EU regulation, ambitious recycling targets were adopted far in excess of industry anticipations, including the reuse and recycling of 85% of cars by 2015. As a result, automotive manufacturers engaged in a virtuous cycle of innovation at design and planning stage. See Pelkmans and Renda, "Does EU regulation hinder or stimulate innovation?". The optimistic tone of the literature on enabling regulation shall however not obscure that firms may follow innovation strategies designed to evade the law. The 2015 Volkswagen NOx (nitrogen oxides) emission scandal highlights that when overly ambitious regulatory targets are adopted, firms have incentives to invest into technologies which game the enforcement system, including malicious software.

[124] Nathalie A. Smuha, "From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence" (*SSRN* December 31, 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3501410>.

### A. Key Concepts

Let us start from the proposition that law and regulation of AI purport to address externalities. By externalities, we mean activities that inflict harm or provide benefits to third parties, ie any party other than the AI system.

The proposition is rooted in mainstream public interest theory. The choice of this framework is not the result of convenience or coincidence, but instead follows the underlying, and often implicit, paradigm of the four models of law and regulation AI discussed above (with the exception, perhaps, of ethics).[125]

Two types of externalities can be distinguished. A negative externality occurs when an AI system imposes costs on third parties. A positive externality appears when an AI system provides benefits on third parties. Positive and negative externalities exist when the AI system (or its governor) fail to internalize or appropriate all or any of those benefits or adverse effects. Economic theory suggests that rational agents overinvest in the supply of activities that produce negative externalities. For example, AI developers may invest in AI systems that reduce the demand for labor and wages, without this being compensated by enough productivity gains that compensate technological unemployment or the delay in the introduction of other productivity enhancing technologies.[126] Conversely, the private sector may underinvest activities in basic and long term AI research and development which yield positive externalities.[127] For example, manufacturers may not invest in ethical standards and “friendly AI” initiatives, because the benefits of this are largely appropriated by third parties. In both configurations, economic theory explains that a public interest-driven government can attempt to correct externalities through the imposition of taxes, the allocation of subsidies or the promulgation of explicit legislative and administrative controls.[128]

Classifying an externality as good or bad involves a good deal of subjective judgment. The replacement of workers by AI systems and machines is a case in point. On the one hand, the externality can be seen as positive, since machines can work more efficiently and faster than people what may lead to a general reduction in prices[129]. On the other hand, it will increase the unemployment rate, and this constitutes a negative externality.

Building on the notion of externalities, we introduce hereafter a novel distinction between three types of externalities. The first type consists in *discrete externalities* (negative of positive). These externalities present the following non-cumulative properties. They are personal, random, rare or endurable. Personal externalities affect third parties at the individual agent level. Random externalities affect all and any third party with equal chance. Rare externalities exhibit low frequency of occurrence. Endurable externalities do not drastically impair the “quality of life” of those who are subject to it or do not radically improve it.[130]

[125] See Daniel Weld and Oren Etzioni, “The first law of robotics (a call to arms)” [1994] 2 AAAI’94: Proceedings of the twelfth national conference on Artificial Intelligence,1042-1047.

[126] Daron Acemoglu and Pascual Restrepo, “Robots and Jobs: Evidence from US Labor Markets” (2020) 128(6) Journal of Political Economy.

[127] Executive office of the President National Science and Technology Council Committee on Technology, “Preparing for the Future of Artificial Intelligence” (October 12, 2016) <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>.

[128] Richard A. Posner, “Theories of Economic Regulation” (1974) 5(2) The Bell Journal of Economics and Management Science, 335-358.

[129] A price drop is not guaranteed though, since the cost of purchasing and controlling robots should be added to the equation.

[130] What Bostrom calls a non pure human enhancement. Nick Bostrom, “Existential risk prevention as global priority” (2013) 4(1) Global Policy, 15-31 (hereafter Bostrom, “Existential risk prevention”).

A typical example of a negative discrete externality is a AI gardener whose visual recognition module dysfunctions and confuses the neighbor's cat with a parasite, ending up spraying the cat with toxic pesticide. A typical example of a discrete positive externality occurs if the AI gardener eradicates parasites when operated at night.

The second type covers *systemic externalities*. They cover third party harm or benefits with the following non-cumulative properties: local, predictable, frequent or unsustainable. By local, we look at harm or benefit that affect a non-trivial segment of the population. By predictable, we envision harm or benefit that is foreseeable for a benevolent authority. By frequent, we consider a repeated occurrence of harm or benefit. By unsustainable, we refer to a non-transitory reduction or increase in well-being of the population class under consideration (given scarce resources). A durable rise in inequalities (poor get poorer, rich get richer) is a case in point.

An often-discussed negative systemic externality consists in the substitution of man by intelligent machines on the factory floor (and the ensuing disappearance of many existing manufacturing jobs, pressure on workers' wages in the long term, etc.). Conversely, a less discussed though equally important positive systemic externality consists in the new complementary jobs that will be created by the introduction of intelligent machines and cognitive computing in industrial sectors (and the corollary reduction in manufacturing costs across the economy as well as transfers of productivity gains to consumers through lower prices).

The third group of externalities comprises existential threats and opportunities created by AIs and robotic applications. To denote their existential nature, we call them as "*existernalities*". Existernalities exhibit several cumulative properties: they are global, improbable, unpredictable and terminal. Global existernalities hit indiscriminately across geographies, demographies and societies. Improbable externalities are those that are usually dismissed by rational wisdom as fictional. Unpredictable externalities are those whose timescale and likelihood of occurrence are improperly assessed. Terminal existernalities have the potential to extinguish humanity as we know it.[131]

We talk of existential properties to denote both biological and philosophical concerns. Existernalities cover "acts which can cause large-scale destruction of lives and property".[132] But they also refer to "acts which can destroy the philosophical and ethical foundations upon which society is built".[133] Germignani advocates pro-active regulation of existernalities.

Negative existernalities include the risk of human extinction,[134] malign superintelligences,[135] lethal autonomous weapons, and other dystopian, terminator-spirited scenarios of machine takeover. Positive existernalities include pure human enhancement,[136] cosmic endowment,[137] virtual immortality, etc. Often, the boundary between a positive and negative existernality is a subjective issue. For instance, time-travel is seen by some as a threat for humanity, and by others as an improvement. Away from science-fiction, Kranzberg talks of dis-benefits and mentions: "advances in medical technology and water and sewage treatment have freed millions of people from disease and plague and have lowered

[131] The concept of "terminality" can be further delineated by distinguishing imminent terminal existernalities that lack the unpredictability property and distant existernalities which fulfill all four properties.

[132] Gemignani, "Laying Down The Law To Robots", 1046.

[133] Ibid.

[134] See Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (Penguin Book, 2006). This also covers dehumanization through the blurring of distinctions between machines and humans.

[135] See Bostrom, "Existential risk prevention".

[136] Pure human enhancement goes beyond the restoration of destroyed human functions. Human enhancement is often opposed to therapy, which "aims to fix something that has gone wrong". But this distinction is not airtight. See Nick Bostrom and Rebecca Roache "Ethical issues in human enhancement" in Jesper Ryberg, Thomas Petersen and Clark Wolf (eds), *New waves in applied ethics* (Pelgrave Macmillan, 2008), 120-152.

[137] See Bostrom, "Existential risk prevention".

infant mortality, these have also brought the possibility of overcrowding the earth and producing, from other causes, human suffering on a vast scale".[138] Gemignani provides an example of a law that seeks to address existernality: "Should there be a law that no machine which carries out a peculiarly human function, such as determining guilt or innocence, be permitted to take a human form?"[139]. Gemignani appears concerned about the existential costs of delegating justice to anthropomorphic machines.

In Table 2 below, we list some examples of discrete and systemic externalities, as well as existernalities.

**Table 2: Typology and Examples of Externalities**

| | | |
|---|---|---|
| Discrete Externality | | Public interest |
| Negative | An industrial robot restarts abruptly and kills a worker on the factory floor. | |
| Positive | Drone spots thief on way to delivery destination, alerts law enforcement which stops the burglar. | |
| Systemic Externality | | |
| Negative | General reduction in privacy across society due to generalized operation of information-hungry AI systems | |
| Positive | Improved disaster responses and humanitarian systems thanks to AI monitoring of population w/o consent | |
| Existernality | | Existential |
| Negative | Permanent state of war following introduction of Lethal Autonomous Weapons ("LAWs") | |
| Positive | Acceleration towards technology frontiers: time-travelling; emulated minds; cosmic exploration | |

Admittedly, this classification is not perfect. A wide spectrum exists between systemic and discrete externalities. Take a malfunctioning self-driving vehicle that drives over a bystander by error and causes serious injuries. Ostensibly, the case does not fall into the "discrete" category because although the case is personal, random, and rare. The impact on the family of the victim as well as the demonstration of a defect in the safety of technology prompt wider social concerns. On the other hand, this illustration cannot be categorized under the "systemic" category since that accident is not frequent, nor local, neither unsustainable despite the fact it caused a decrease in the standard-of-living of the injured civilian.

### *B. Normative Implications*

The normative implications from the above conceptual framework follow a logical progression from existing law to the development of public policy.

The resolution of discrete externalities should be left to existing laws. Society defers to the decentralized courts system which will process discrete externalities on a case-by-case basis. Disputes are solved *ex post* through the application of the general rules of property, contract and liability and other specific laws. This is acceptable because discrete externalities cannot affect society by any

[138] Kranzberg, "Kranzberg's Laws", 547.

[139] Gemignani, "Laying Down The Law To Robots", 1050.

significant order of magnitude. Moreover, this regulatory approach is efficient, because it allows a degree of decisional experimentation, benchmarking, and cross-fertilization.

When more severe threshold effects are encountered with systemic externalities, society should contemplate public policy development. The question is whether *ad hoc* law or regulation ought to be adopted to correct the systemic externality. Here are some examples of such questions in relation to negative externalities: must a specific tax be introduced in automation-intensive industries subject to creative destruction?; must black-box[140] requirements be imposed on manufacturers of AI systems confronted with moral dilemmas like the trolley problem?; must specific privacy regulation be adopted on the second-hand AI systems market to protect data subjects, including previous governors? Likewise, examples abound for positive systemic externalities: given the public goods nature of infrastructure and collective action problems amongst competing producers, must subsidies be allocated for the construction of controlled environments for AI systems (for example, specific road infrastructure for driverless cars)?; should developers and manufacturers of generative AI technologies enjoy statutory immunity for damages caused by their inventions?[141]; Should intellectual property regimes be relaxed to enable open, transparent and peer-scrutinized research processes in AI systems with the goal of friendly AI?

Regulatory responses to systemic externalities must be subject to *ex ante* and *ex post* impact assessment. By *ex ante* impact assessment, we refer to the prospective cost-benefit evaluation of future regulatory options. By *ex post* impact assessment, we consider retrospective cost-benefit measurement of experimented regulatory options. In both cases, society experiments various regulatory options in dedicated zones of the real-life environment, and proceeds to evaluate the results of such tests. In Japan, for instance, the creation of so-called "Tokku zones" system has entitled robot manufacturers to conduct practical tests on public roads and environments.[142] This mixed *ex ante* and *ex post* approach limits risks of Collingridge type quandaries and reduces risks of disabling regulation.

Existernalities create concerns of such levels that they can be *ex ante* subject to law and regulation, without prior AI and robotic experimentation, implementation or realization. Given their global nature, the regulation of existernalities should tentatively be decided by international organizations. However, international organizations are often paralyzed by gridlock on existential issues (like peacekeeping or climate change) due to their wide membership. In the AI field, endless discussions have taken place at the United Nations over a proposed ban on the use of lethal autonomous weapons ("LAWs"). Regional institutions (like the EU) might be better forums for the initial regulation of existernalities. And yet, this is not a given. For example, the EU HLEG on AI failed to adopt red lines of research on AI consciousness, LAWs or citizen scoring systems.

In addition, the fact that existernalities are "black swans" implies a degree of fatality in terms of our failure to anticipate them.[143] Conversely, knee-jerk regulatory responses cannot be excluded in democratic systems. Overall, a degree of expert and technocratic input in decision-making is therefore appropriate. In this context, the involvement of standard-setting organisations (like the IEEE, SAE, the ISO and many others) may play a useful contributive role to the definition of early positions on existernalities. Last, objections to the costs of prohibitive *ex ante* intervention are not material, because the costs of type II errors (false negatives) in relation to existernalities are higher than the costs of type

[140] A black-box algorithm is an algorithm for which it is impossible to know how it reach a particular output from given inputs. According the High-Level Expert Groups on Artificial Intelligence, black-box requirements should be the implementation of explicability, traceability and auditability. See AI HLEG Guidelines.

[141] For a precedent, see Senate Bill: S. 1458 (103rd) General Aviation Revitalization Act of 1994, known as GARA.

[142] Yueh-Hsuan Weng, Yusuke Sugahara, Kenji Hashimoto and Atsuo Takanishi "Intersection of 'Tokku' special zone, robots, and the law: a case study on legal impacts to humanoid robots" (2015) 7(5) International Journal of Social Robotics, 841-857.

[143] Events that come as surprises to most, if not all. See Taleb, *Black Swan*.

I errors (false positives). A type II error occurs when we fail to remedy a serious existential risk in probability and/or intensity terms. A type I error occurs when we wrongly remedy a moot existential risk in probability and/or intensity terms. Immediately one understands that the cost of a type II error is existential, whilst this is not necessarily the case for a type I error. The cost of the latter is thus more acceptable than the cost of any type II error which will always be existential. But there is more. A type II error in relation to existernalities is not reversible, because humanity has disappeared. This excuses any and every type I error in relation to existernalities.

## Conclusion

This paper has attempted to describe models of law and regulation for AI. Its main ambition is primarily descriptive: help readers make sense of developing legal frameworks in this ever evolving, and quite anarchic, area of the law.

In addition, this paper has developed a normative case for a new model of law and regulation for AI. The model proposes to index the intensity of regulatory response upon the nature of the externality created by an AI application. When AI-generated externalities are discrete, societies should defer to *ex post* litigation before courts. When AI-generated externalities are systemic, societies planners should envision *ex ante* regulation, but carefully test and experiment. This meshes the benefits of anticipation and empiricism and avoid Collingridge dilemma as well as disabling regulation problems. Last, when AI-generated externalities are existential, societies should consider *ex ante* intervention, and bring into it a degree of expert deliberation.

Our proposed model is not only about distinguishing levels of regulatory response based on a probabilistic reasoning and classification of externalities regarding their frequency, severity, and globality. It also overcomes the pitfall of the “tyranny of numbers” and the “aura of precision” by making ethical concerns central.[144]

[144] Bob Heymand and Mike Titterton, “Introduction” in Bob Heyman, Andy Alaszewski, Monica Shaw and Mike Titterton (eds), *Risk, Safety and Clinical Practice: Health care through the lens of risk* (Oxford University Press 2008) 3 and 11.

**Author contacts:**

**Nicolas Petit**

Robert Schuman Centre for Advanced Studies, European University Institute
Villa Schifanoia, Via Boccaccio 121
I-50133 Florence

Email: Nicolas.petit@eui.eu

**Jerome De Cooman**
University of Liege (ULiege) and Liege Competition and Innovation Institute (LCII)
Quartier Agora, Place des Orateurs 1
B-4000 Liege

Email: Jerome.decooman@uliege.be

With the support of the Erasmus+ Programme of the European Union

The European Commission supports the EUI through the European Union budget. This publication reflects the views only of the author(s), and the Commission cannot be held responsible for any use which may be made of the information contained therein.