Revista de la Facultad de Derecho de México
Tomo LXXI, Número 279, Enero-Abril 2021
DOI: http:// dx.doi.org/10.22201/fder.24488933e.2021.279-1.78811

Hasta la vista, baby: reflections on the risks of algocracy, killer robots, and artificial superintelligence**

Hasta la vista, baby: reflexiones sobre los riesgos de algocracia, robots asesinos y superinteligencia artificial

Pedro Rubim Borges Fortes[*]

Abstract: The neologism Algocracy may mean government or governance by algorithms. Architects of artificial intelligence have perspectives on killer robots and government by artificial superintelligence and are engaged in public debate on both themes. The risks of being dominated by artificial superintelligence and of being subjected to undemocratic, unconstitutional or illegal algo norms inspires our reflection. Institutions should organize rules of the game that prevent machine learning algorithms from learning how to dominate humans. Algorithms need new design requirements to incorporate responsibility, transparency, auditability, incorruptibility, and predictability. The algorithmic responsibility of the state, national public policies for developing a trustworthy AI, and the

[*] DPHIL (Oxford), JSM (Stanford), LLM (Harvard), MBE (Coppe-UFRJ), BA (PUC-Rio), LLB (UFRJ). Visiting Professor at the Doctoral Program of the National Law School at UFRJ. Chair of the Working Group Law and Development at the Research Committee of Sociology of Law. Chair of the Collaborative Research Network Law and Development at the Law and Society Association. Convenor of the stream Exploring Legal Borderlands: Empirical and Interdisciplinary Approaches at the Socio-Legal Studies Association. International Director of the Brazilian Institute for Studies of Tort Law (IBERC). Research Associate at the Laboratory of Institutional Studies (LETACI). ORCID ID: 0000-0003-0548-4537. Contacto: <pfortes@alumni.stanford.edu>. Fecha de recepción: 15/11/2020. Fecha de aprobación: 18/02/2021.

algorithmic law of killer robots and artificial superintelligence could reduce the risks of algocracy. The particular character of algorithms demands a special discipline to control their power, architecture, and commands. law and government can channel the development and use of killer robots, eventually even setting a global prohibition of autonomous weapons. Likewise, the threat of government by algorithms posed by the emergence of an artificial superintelligence that dominates humankind also requires the development of a new algorithmic law that establishes checks and balances and controls the technological system.

Keywords: Algorithmic Law; Artificial Intelligence; Algocracy; Superintelligence; Killer robots.

Resumen: El neologismo Algocracia puede significar gobierno o gobernanza por algoritmos. Los arquitectos de la inteligencia artificial tienen perspectivas sobre los robots asesinos y el gobierno mediante la superinteligencia artificial y participan en el debate público sobre ambos temas. El riesgo de ser dominado por una superinteligencia artificial y de ser sometido a normas algo antidemocráticas, inconstitucionales o ilegales inspira nuestra reflexión. Las instituciones deben organizar reglas del juego que eviten que los algoritmos de aprendizaje automático aprendan a dominar a los humanos. Los algoritmos necesitan nuevos requisitos de diseño para incorporar responsabilidad, transparencia, auditabilidad, incorruptibilidad y previsibilidad. La responsabilidad algorítmica del estado, las políticas públicas nacionales para desarrollar una IA confiable y la ley algorítmica de los robots asesinos y la superinteligencia artificial podrían reducir los riesgos de la algocracia. El carácter particular de los algoritmos exige una disciplina especial para controlar su poder, arquitectura y comandos. la ley y el gobierno pueden canalizar el desarrollo y uso de robots asesinos, y eventualmente incluso establecer una prohibición global de armas autónomas. Asimismo, la amenaza de gobierno por algoritmos que plantea la aparición de una superinteligencia artificial que domina a la humanidad también requiere el desarrollo de una nueva ley algorítmica que establezca frenos y contrapesos y controle el sistema tecnológico.

Palabras clave: Ley algorítmica; Inteligencia artificial; Algocracia; Superinteligencia; Robots asesinos.

[**] This article originated from a research agenda on the mathematical turn on legal analysis and the normativity embedded in algorithms in 2014 and I had the opportunity to present papers and discuss ideas with several colleagues at various events. I would like to thank particularly David Restrepo-Amariles, Benoît Frydman, Gregory Lewkowicz, Arnaud van Waeyenberge, Delphine Dogot, Krishna Gummadi, Hakan Hydén, Ji Weidong, Stefan Larsson, Alfons Bora, Gunther Teubner, Thomas Vesting, Ricardo Campos, Mavis Maclean, Ioannis Kampourakis, Susan Sterett, Laurence Dumoulin, Sabine Gless, Lawrence Friedman, William Twining, João Falcão, Pedro Delfino, Gulherme Magalhães Martins, Nelson Rosenvald, Ricardo Morishita, Edson Alvisi, Eduardo Tostes, Ricardo Cueva, Roberto Freitas, Luiz Roberto Ayoub, Plínio Lacerda, and Carlos Bolonha for these opportunities of exchange and feedback that contributed to the refinement of reflections made here. I must also greet Michael Asimow, Christine Corcos, Peter Robson, Kathryn Brown and all colleagues from the law and popular culture community with a ‘Hasta La Vista’. Likewise, I am grateful to Dr. Jorge Fernández Ruiz, Manuel de J. Jiménez Moreno, Ms. María del Carmen Maya Flores, and all the editorial team of the *Revista de la Facultad de Derecho de Mèxico* for their invaluable support with this publication. Errors are mine.

DOI: http:// dx.doi.org/10.22201/fder.24488933e.2021.279-1.78811

I. Introduction

"Hasta la vista, Baby" is the well-known catchphrase associated with Arnold Schwarzenegger as part of his role as a robot in the blockbuster film *Terminator 2: Judgment Day* (1991). Initially, the phrase provides evidence of the learning potential of artificial intelligence (AI), as the cyborg learns new expressions from humans and quickly improves his capacity of communicating. Later, the phrase was also used in the film to mark the elimination of an opponent by the terminator. The idea that robots may become super intelligent and eventually also terminators may inspire our reflection on how to reduce the risks posed by *algocracy*.

Importantly, popular culture often provokes innovative thinking on law.[1] In her study of representation of robots in science fiction, for instance, Christine Corcos identified claims to self-recognition that could eventually lead to their legal personhood and protection of their fundamental rights.[2] Her research anticipated a discussion on the potential development of specific civil law rules on robotics that could grant legal personality to AI and to attribute responsibilities, duties, and rights to robots.[3] This perspective

[1] Friedman, Lawrence M. "Law, lawyers, and popular culture." *The Yale Law Journal* 98.8 (1989): 1579-1606; Asimow, Michael, Brown, Kathryn, and Papke, David, (eds.), *Law and popular culture: International perspectives*, Cambridge Scholars Publishing, 2014; Greenfield, Steve, Guy Osborn, and Peter Robson, *Film and the law: The cinema of justice*. Bloomsbury Publishing, 2010; Twining, William. "Law and Literature: A Dilettante's Dream?.", *Journal of the Oxford Centre for Socio-Legal Studies*, n. 2, 2017, pp. 126-139.

[2] Corcos, Christine A., "More Human Than Human: How Some SF Presents AI's Claims to the Right to Life and Self-Determination", *Oxford Journal of Socio-Economic Studies, Hilary Term*, 2017.

[3] Turner, Jacob, "Legal personality for AI."*Robot Rules*. Palgrave Macmillan, Cham, 2019, pp. 173-205; Schirmer, Jan-Erik, "Artificial Intelligence and Legal Personality: Introducing "Teilrechtsfähigkeit": A Partial Legal Sta-

seems to be more integrated with the rise of AI by examining ways of accommodating robots within contemporary human societies.[4]

In contrast to these integrated perspectives, apocalyptical visions of the future of artificial intelligence emulate dystopian narratives, like the one depicted in the "Terminator" series, in which an AI system called Skynet controls nuclear missiles and initiate a plan to exterminate humanity from the planet. Only a fictional narrative thirty years ago, some of the architects of AI consider that there is a concrete possibility that technology could misuse itself when it becomes superintelligent and pursue different objectives from what humans really want.[5]

One of the greatest risks posed by this setting of artificial superintelligence could be the emergence of *algocracy*. Originally coined by Indian sociologist A. Aneesh, *algocracy* was defined as a code-based governance system consisting of programming schemes embedded in global software platforms that control performance, guide action, and contribute to decision-making.[6] In a nutshell, *algocracy* implies the "rule of the algorithm" instead of the rule of law[7] and an exercise of political authority that legitimizes itself by the routes programmed and embedded in the underlying computer code.[8] Algocratic governance, according to

tus Made in Germany."*Regulating artificial intelligence*. Springer, Cham, 2020, pp. 123-142; Van Genderen, Robert van den Hoven, "Do we need new legal personhood in the age of robots and AI?." *Robotics, AI and the Future of Law*. Springer, Singapore, 2018, pp. 15-55; Gordon, John-Stewart, "Artificial moral and legal personhood", *AI & SOCIETY*, 2020, pp. 1-15.

[4] On the distinction between the integrated and the apocalyptical perspectives, see Eco, Umberto, *Apocalittici e integrati*, vol. 27, T. Bompiani, 1984.

[5] Ford, Martin, *Architects of Intelligence: The truth about AI from the people building it*, Packt Publishing Ltd, 2018: 98.

[6] Aneesh, Aneesh, "Global labor: Algocratic modes of organization.", *Sociological Theory*, 27.4, 2009, p. 349.

[7] *Idem*, 350.

[8] *Idem*, 356.

Aneesh, is coded as a program that determines the range of possible action automatically.[9]

As any neologism, *algocracy* may be interpreted in different ways. John Danaher, for instance, identified a narrow definition –a system in which AI seizes control of governmental decision-making bodies and exercise power according to its needs and interests and a broader one– a governance system organized and structured on the basis of computer program algorithms.[10] In his analysis, Danaher decided to focus on *algocracy* as governance by algorithms instead of a government by algorithms, but both conceptions deserve our analysis.[11]

The originality of the present article comes exactly from the fact that it reflects on *algocracy* from both perspectives –as government by algorithms and governance by algorithms. I will argue that institutions should organize rules of the game that prevent machine learning algorithms from learning how to dominate humans. I will also argue that we should set clear, fair, and proportionate guidelines for designing algorithms that influence public policies, shape decision-making processes, and affect social interests.

This article will pursue these arguments, by reflecting on the case studies of killer robots and of artificial superintelligence. Inspired by popular culture and exemplified by scenes of *Terminator 2: Judgment Day* (1991), both cases entered in the public debate of the architects of AI and have mobilized calls for reflection on the ethical values of contemporary technology. The relevance of this article comes from the growing impact of the normativity embedded in algorithmic formulas over society.

Additionally to this introduction, section two will review the ethical challenges related to killer robots and to government by artificial superintelligence based on the perspectives of some rele-

[9] *Idem.*

[10] Danaher, John, "The threat of algocracy: Reality, resistance and accommodation", *Philosophy & Technology*, 29.3, 2016, p. 249.

[11] *Idem.*

vant actors in academia and corporations, leading to the conclusion that algorithms need new design requirements and superintelligence requires new ethical standards. Section three explores legal responses to reduce the risks posed by *algocracy*, examining the algorithmic responsibility of the state, national public policies for developing a trustworthy AI, and the limits and possibilities of the algorithmic law of killer robots and artificial superintelligence. Section four will bring concluding remarks.

II. Ethical Challenges Related to Killer Robots and Government by Artificial Superintelligence According to the Architects of AI

In a recent book, Martin Ford published interviews conducted with twenty-three prominent people in the AI scene. His book *Architects of Intelligence: The Truth about AI from the people building it* provides excellent material for anyone interested in learning more about the ethical challenges posed by AI and it also brings empirical evidence on their perspectives on killer drones and government by artificial superintelligence.[12] In this section, I refer to these interviews to map these ethical challenges according to a representative sample of prominent leaders of this technological community.

Yoshua Bengio, for instance, has been very active against killer robots and signed a letter aimed at the Korean Advanced Institute of Science and Technology (KAIST) to prevent the development of military robots without a human in the loop.[13] According to the Scientific Director of the Montreal Institute for Learning Algorithms, difficult moral questions should never be put in the hands of machines, because current AI – and the AI that we can

[12] Ford, Martin, *Architects of Intelligence: The truth about AI from the people building it*, Packt Publishing Ltd, 2018.

[13] *Ibidem,* p. 31.

foresee in the near future – have no moral understanding of what is right and what is wrong.[14] In his opinion, however, the existential threat from super intelligent AI is not a concern nowadays, because these scenarios are not realistic and not compatible with how AI is built right now.[15]

Stuart Russel also focused on the potential risks of weaponized AI, expressing concern that autonomous weapons may lead to a new arms race and that power over life and death should not be handled to a machine to decide.[16] Interestingly, the Professor of Computer Science from UC Berkeley recorded a short film called *Slaughterbots* to raise awareness to the public that autonomous weapons are no longer science fiction restricted to our imagination of Skynet and Terminators and that AI warfare technologies are feasible today.[17] Stuart Russel is also extremely concerned with the potential risks of machines with dominant effect on the real world, because intelligence represents power over the world and something with greater intelligence would also have more power.[18] His cautionary note is that AI needs to remain under human control and should retain the property of corrigibility, being able to be corrected and eventually to be switched off.[19]

Nick Bostrom warns about the risks posed by the creation of an artificial agent capable of achieving its own objectives due to its superintelligence, that would optimize goals that are contrary to our human values and eventually win.[20] Even if he doesn't see need for regulations related to machine superintelligence now, the Director of the Future of Humanity Institute considers that we

[14] *Ibidem,* pp. 31-32.

[15] *Ibidem,* p. 33.

[16] *Ibidem,* pp. 58-59.

[17] *Ibidem,* p. 60-61.

[18] *Ibidem,* p. 62.

[19] *Ibidem,* pp. 66-67.

[20] *Ibidem,* p. 98.

need to discuss over values and how different values should guide the use of this technology.[21]

On the other hand, Yann Lecun seems skeptical about these issues, considering that militaries are going to use AI technology for surgical actions that will look more like police operations and less like weapons of mass destruction.[22] Likewise, the Chief AI Scientist of Facebook states that "we should not worry about the Terminator scenario", because he is skeptical that we would create out-of-control human-level intelligence that would take over the world.[23]

Demis Hassabis consider that he has a more nuanced view of these ethical issues and that his view is in the middle of the more extreme perspectives, based on the opinion that the technology itself is neutral and depends on human design and decisions regarding use and distribution of benefits.[24] Their premise that AI should remain under meaningful human control and be used for socially beneficial purposes implies their support for banning autonomous weapons, because a meaningful level of human judgment and control is necessary to guarantee that weapons are used in ways that are necessary and proportionate.[25]

Andrew Ng considers the debate on artificial superintelligence so premature that he states that "worrying about AGI evil killer robots today is like worrying about overpopulation on the planet Mars".[26] Likewise, the CEO of AI Fund considers that any new technology –internal combustion engine, electricity, and integrated circuits– may be useful for the military and that the same is true for AI.[27]

[21] *Ibidem,* pp. 101-102.

[22] *Ibidem,* p. 138.

[23] *Ibidem,* pp. 135-136.

[24] *Ibidem,* p. 177.

[25] *Ibidem,* p. 179.

[26] *Ibidem,* p. 202.

[27] *Ibidem,* p. 203.

Rana El Kaliouby doesn't subscribe to the fears that robots are going to take over humanity, because humans are designing these systems, defining their deployment, and can turn the switch off.[28]

Barbara Grosz supports teaching students to design more ethical programs, reminding that it is very easy to put something very bad on a drone.[29] But the Higgins Professor of Natural Sciences at Harvard University considers extreme the position to stop working with AI, especially because of all the wonderful ways in which AI can improve the world and make it a better place.[30]

Judea Pearl warns that we have to worry about artificial intelligence, to understand what we build and that we are breeding a new species of intelligent animals that are initially domesticated but eventually will assume their own agency.[31] In the opinion of the Professor of Computer Science and Statistics of UCLA, "we should absolutely be cautious about the possibility that we are creating a new species of super-animals, or in the best case, a species of useful, but exploitable human beings that do not demand legal rights or minimum wage".[32]

Jeffrey Dean provides also a cautious tone on the challenges related to the development of AGI, stating that it should be done ethically and based on sound decision-making.[33] The Head of AI at Google referred to their AI principles document as a clear guideline on how to approach problems, tackle with these approaches, what will not be done with these sorts of issues.[34]

While examining the challenges related to weaponization of drones, Daphne Koller acknowledges the existence of security risks to AI systems, but she puts it that she doesn't know

[28] *Ibidem,* p. 221.

[29] *Ibidem,* p. 350.

[30] *Ibidem,* pp. 350-351.

[31] *Ibidem,* p. 371.

[32] *Idem.*

[33] *Ibidem,* p. 384.

[34] *Idem.*

"that they're qualitatively different to the same risks with older technologies".[35] After being pressed on this subject, the CEO of Insitro agrees that technological development increased the ability to kill larger numbers of people, but she still "wouldn't say that stories of intelligent killer drones are more dangerous than someone synthesizing a version of smallpox and letting it loose".[36] Likewise, she thinks that the discussion on the control problem of superintelligent agents that might set their own goals and implement them in harmful ways is premature, because there are several breakthroughs that need to happen and superintelligence is not going to be an emergent phenomenon, but rather an engineered system.[37]

David Ferrucci considers that there is cause for concern anytime you give leverage to a machine, putting it in control over something that can amplify an error or the effect of a bad actor and lead to a significant disaster.[38] However, the Elemental Cognition Director of Applied AI at Bridgewater Associates is less concerned about the possibility that the machine might develop its own goals and lay waste to the human race, because one would have to program the computer to do something like that and there are fewer incentives for machines to react like that.[39]

Rodney Brooks worries less about a self-aware AI doing something willful or bad and more about human actors exploiting the weaknesses of these digital technologies to do bad things.[40] The Chairman of Rethink Robotics thinks that the weaponization of robots and drones is very possible today, but doesn't think that keeping AI out of the military is a solution and that instead we

[35] *Ibidem,* p. 398.

[36] *Ibidem,* pp. 399-400.

[37] *Ibidem,* p. 400.

[38] *Ibidem,* p., 417.

[39] *Idem.*

[40] *Ibidem,* p. 439.

should legislate against what we don't want to happen.[41] Regarding the control problem of superintelligence, Rodney Brooks thinks that we have no clue about the future of AI and that isolated academics living in a bubble away from the real world play just a power game, but we won't know what these technologies will look like before they arrive.[42]

Cynthia Breazel also is less concerned about superintelligence enslaving humanity than about people using these technologies to do harm.[43] Not only she thinks that superintelligence would not emulate the evolutionary forces that drove the creation of human motivation and drives, but also the enormous concentration of talent, funding, and people for the creation of superintelligence has not yet been mobilized within academia, governments, and corporations.[44] On the other hand, the MIT Media Laboratory Founder sees real risks around autonomous weapons.[45]

Joshua Tenenbaum finds reasonable that some people are thinking about the risks posed by superintelligence and that we could imagine some kind of superintelligence could pose an existential risk to humanity, but believes that other existential risks are much more urgent, like the understanding of moral principles and AI value alignment.[46] Because he thinks that the idea that machines would decide for themselves to take over the world is so remote, the Professor of Computational Cognitive Science at MIT is more concerned about the risks related to the development of increasingly powerful algorithms that may support selfish purposes and for evil or bad deeds.[47]

[41] *Ibidem,* p. 440.

[42] *Ibidem,* p. 440-441.

[43] *Ibidem,* p. 456-457.

[44] *Ibidem,* p. 457-458.

[45] *Ibidem,* p. 457.

[46] *Ibidem,* p. 488.

[47] *Ibidem,* pp. 484-485.

Oren Etzioni spontaneously pointed to autonomous weapons as a big concern and a “scary proposition, particularly the ones that can make life-or-death decisions on their own”.[48] In his opinion, however, the focus of concern should be on the autonomy to make life-and-death decisions on their own which is something that we can choose as society to meter out, but intelligence could actually help save more lives, by getting these weapons more targeted or by getting them abort when the human cost is unacceptable.[49] For the CEO of the Allen Institute for Artificial Intelligence, however, the threat of artificial superintelligence is more a subject for contemplation of a small number of philosophers rather than a subject for general concern and any practical action at this moment.[50]

Bryan Johnson compared and contrasted the concerns on artificial superintelligence made by Nick Bostrom and by Elon Musk, praising the way Nick Bostrom initiated and framed the whole discussion, but criticizing Elon Musk for creating and inflicting fear among the general public.[51] Interestingly, the interviewer Martin Ford asked the question on the concerns related to superintelligence to the interviewees often referring to Nick Bostrom and Elon Musk as references to this existential threat.

As Nick Bostrom positioned himself as a clear reference in this discussion according to the architects of AI, we should summarize his main concerns about it. First, AI algorithms must be *transparent to inspection*, especially when they perform cognitive work with social dimensions. Second, AI algorithms must be *predictable to those they govern*, providing stability to the social environment within which citizens may optimize their own lives, just like system does with the predictability of judicial precedents. Third, AI algorithms must be *robust against manipulation*, so that

[48] *Ibidem,* p. 506.

[49] *Ibidem,* p. 507.

[50] *Ibidem,* p. 506.

[51] *Ibidem,* p. 523.

it guarantees information security. Fourth, the ethical discussion should also include the capacity to *attribute blame for the person responsible* for getting something done.[52]

Constructing a trustworthy AGI will require different methods and way of thinking, such as an AGI that thinks like an engineer concerned about ethics and not just a simple product of ethical engineering. In this setting, verifying the safety of the system becomes a greater challenge, because rather than verifying the system's safe behavior in all operating contexts, we must verify what the system is trying to do, as the local and specific behavior of AGI is not predictable.[53]

Particularly the problem of superintelligence consists of a sufficiently intelligent AI that could understand its design and could redesign itself to become even more intelligent in a positive feedback cycle that could lead to an intelligent explosion. The stakes are no longer individual ones, but rather global as humanity could be extinguished and replaced. On one hand, intelligence seems impossible to control and control over the initial programming may not translate into influence on its later effect on the world. On the other hand, human civilizations exhibit directional change in the sense that our ethical values evolve and the discipline of machine ethics must commit itself to seek human-superior niceness.[54]

In summary, AI algorithms need new design requirements to incorporate responsibility, transparency, auditability, incorruptibility, and predictability. Superintelligence presents us with the challenge of stating an algorithm that outputs superethical behavior.[55] Therefore, these ethical challenges invite also our reflection

[52] Cfr. Bostrom, Nick and Yudkowsky, Eliezer, *Ethics of Artificial Intelligence*, in Ramsey, William and Frankish, Keith (eds.), Cambridge Handbook of Artificial Intelligence, CUP, 2011.

[53] *Idem.*

[54] *Idem.*

[55] *Idem.*

on potential legal responses that incorporate ethics into algorithmic formulas and reduce risks posed by killer drones, government by artificial superintelligence, and *algocracy*.

III. Algorithmic Law as the Response to Risks of *Algocracy*, Killer Robots, and Artificial Superintelligence

The case studies of killer robots and government by superintelligence are paradigmatic for our analysis of the potential legal responses to reduce the risks posed by *algocracy*, because they provide prodigious examples of the existential threats that algorithms may bring to fundamental rights, democracy, and the rule of law. The mere existence of autonomous weapons with decision-making capacity to kill individuals without human overview, intervention or any form of control would challenge the foundations of humanitarian law and human rights. For instance, who should be deemed responsible for a war crime attributed to a killer robot, the programmer, the commander or the machine itself?[56] Likewise, a system of artificial superintelligence that seizes control of governmental decision-making bodies and exercises power according to its needs and interests would be functionally equivalent to a dictatorship that concentrates power and exercises it in a authoritarian way.

Importantly, there is a growing awareness of the state algorithmic responsibility as a duty to protect citizens and consumers from violations of diffuse, collective and individual rights resulting from technological wrongdoings. In this context, states have to develop technological capacity to conduct algorithmic audits of private corporations and to establish effective algorithmic regula-

[56] Sparrow, Robert, "Killer Robots", *Journal of Applied Philosophy*, vol. 27, n. 1, 2007, pp. 69-73.

tion with accountability.[57] Institutional intervention is necessary for the protection of collective rights, prevening that asymmetries of power and information cause harmful effects on digital consumers.[58]

State institutions should regulate algorithmic formulas that may negatively impact coordination, redistribution, and deliberation on legally protected interests.[59] Part of automated decision-making processes,[60] algorithmic codes contain embedded normativity[61] and are positioned in the borderlands of law and technology.[62] In summary, the state's algorithmic responsibility means that rights must be defended and state actors need to deve-

[57] Borges Fortes, Pedro Rubim, Magalhães Martins, Guilherme and Farias Oliveira, Pedro, *A Case Study of Digital Geodiscrimination: How Algorithms May Discriminate Based on the Geographical Location of Consumers*, Droit et Société, forthcoming.

[58] Borges Fortes, Pedro Rubim, *Responsabilidade Algorítmica do Estado: Como as Instituições Devem Proteger Direitos dos Usuários nas Sociedades Digitais?*, in Magalhães Martins, Guilherme, and Rosenvald, Nelson, (eds), *Responsabilidade Civil e Novas Tecnologias,* Indaiatuba, Foco, 2020.

[59] *Idem.*

[60] Ferguson, Andrew Guthrie, *The rise of big data policing: surveillance, race, and the future of law enforcement*, New York, NYU, 2017; Virginia Eubanks, *Automating inequality: how high-tech tools profile, police, and punish the poor,* New York, St Martin's Press, 2017; Umoja Noble, Safiya, *Algorithms of oppression: how search engines reinforce racism*. New York, NYU, 2018.

[61] Lawrence Lessig, *Code and Other Laws of Cyberspace*, New York, Basic Books, 1999.

[62] Borges Fortes, Pedro Rubim and Kampourakis, Ioannis, „Exploring Legal Borderlands: Introducing the Theme."*REI-Revista Estudos Institucionais*, 5.2, 2019, pp. 639-655; Borges Fortes, Pedro Rubim, "An Explorer of Legal Borderlands: A Review of William Twining's Jurist in Context", A Memoir, *REI-Revista Estudos Institucionais,* 5.2, 2019, pp. 777-790.

lop expertise and capacity to exercise periodic review and supervision of these algorithms.[63]

In terms of public policy, governments also need to guarantee that artificial intelligence remains trustworthy and nowadays states are encouraged to develop their own national strategies to secure technological development.[64] In recent years, for instance, the EU authorities released three seminal documents on a trustworthy AI: The White Paper on Artificial Intelligence;[65] The Policy and Investment Recommendations for a Trustworthy AI;[66] and the Ethics Guidelines for Trustworthy AI.[67] Among these EU guidelines are ethical principles of *respect for human autonomy*, *prevention of harm*, *fairness*, and *explicability* which somehow echo the recommendations made by Nick Bostrom on the ethics of AI.[68] National AI policies should include computational ethics and require that algorithms incorporate responsibility, transpa-

[63] Borges Fortes, Pedro Rubim, Magalhães Martins, Guilherme and Farias Oliveira, Pedro, *A Case Study of Digital Geodiscrimination: How Algorithms May Discriminate Based on the Geographical Location of Consumers*. Droit et Société, forthcoming.

[64] Larsson, Stefan, Ingram Bogusz, Claire and Schwarz, Jonas Andersson (eds.), *Human-Centred AI in the EU: Trustworthiness as a Strategic Priority in the European Member States*, Elf, 2020.

[65] Consultado en: <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>.

[66] Consultado en: <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>.

[67] Consultado en: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

[68] I had the opportunity to discuss recently these ethical challenges in my analysis of the Portuguese national AI policy. See Borges Fortes, Rubim, Pedro, *AI Policy in Portugal: Ambitious Yet Laconic About Legal Routes Towards Trustworthy AI*, in Larsson, Stefan, Bogusz, Claire Ingram, and Andersson Schwarz, Jonas (eds.), *Human-Centred AI in the EU: Trustworthiness as a Strategic Priority in the European Member States*, Elf, 2020.

rency, auditability, incorruptibility, and predictability within their new design requirements.

The case study of killer robots provides a prodigious example on the possibilities and limits of algorithmic law. In his study of the threat of *algocracy*, Danaher decided not to analyze the possibility of superintelligent AI controlling the world, explaining that it may happen in the future, but that he is more concerned with mundane problems related to governance.[69] Danaher defined *algocracy* as "a system in which algorithms are used to collect, collate, and organize the data upon which decisions are typically made and to assist in how that data is processed and communicated through the relevant governance system".[70]

In his opinion, the experience with military drones generated useful distinctions between types of robotic weapon system which are relevant for the reflection on entirely automated systems and the role of humans in reviewing and scrutinizing the recommendations made by algorithms: *human-in-the-loop weapons* – robots can only select targets and deliver force with a human command; *human-on-the-loop weapons* – robots can select targets and deliver force on their own, but there is human oversight and the possibility of human override; *human-out-of-the-loop weapons*: robots act autonomously, selecting targets, and delivering force without human oversight or override.[71]

Among the concerns emerging from *algocracy*, data is collected and used in a covert and hidden manner without consent of data owners (*hiddenness concern*) and their operations are inaccessible or opaque to human reason and understanding (*opacity concern*).[72] Reliance on algocratic systems limits the scope for active human participation in and comprehension of decision-making

[69] Danaher, John, "The threat of algocracy: Reality, resistance and accommodation", *Philosophy & Technology*, 29.3, 2016, pp. 246-247.

[70] *Ibidem,* p. 247.

[71] *Ibidem,* p. 248.

[72] *Ibidem,* p. 249.

procedures, threatening their legitimacy.[73] Danaher provides a list of solutions to the threat of algocracy, such as: (1) insisting upon human review of algorithms; (2) enhancing knowledge of human beings; (3) embracing technologies for radical transparency; (4) establishing partnerships between individuals and algorithms.[74]

In his discussion of these potential solutions to the threat of algocracy, however, Danaher highlights their limitations, by pointing out that (1) regulatory overview of algorithms and procedural data due process may be insufficient because the possibility for human review may be blocked for non-interpretable data-mining processes and complex operations within a broader ecosystem of connected algorithms; (2) algocratic systems are likely to rely on processes and capacities that are radically beyond what is possible for human beings to understand and any strategy to enhance knowledge would probably lead to the emergence of a group of epistemically elite human beings; (3) technologies for radical transparency do not correct asymmetries of knowledge, power, and even of information, as the rational basis for decision-making is not defined through collection and processing of data by humans, but rather by a complex ecosystem of algorithms; (4) partnerships between individuals and algorithms would be limited not only by the epistemic elitism, but also by the fact that algorithms are designed by powerful corporations (companies, governments, and universities) and not by individual citizens, who end up as passive recipients of the wisdom of their AI assistants and not true agents involved in understanding and shaping our destinies.[75]

Our reflection of the problems posed by *algocracy* require that we also investigate the normativity embedded in algorithms. Further developing Lessig's insight that "code is law"[76], Hydén states

[73] *Ibidem,* p. 254.

[74] *Ibidem,* pp. 258-265.

[75] *Idem.*

[76] Lessig, Lawrence, *Code: And other laws of cyberspace*, ReadHowYou-Want. Com (2009).

that correctly that algorithms are norms, as these technical formulas contain "strong inherent normativity".[77] Interestingly, however, because algorithms encodes social values into the digital architecture of digital platforms that surrounds us, we need to develop the capacity to see and comprehend the algorithmic normativity emanating from technological systems and to perceive these norms as concrete phenomena of contemporary digital societies.[78]

Therefore, Hydén coined the term *algo norms*, as a conceptual strategy to distinguish and to separate the technological dimension of the technical instruction from the normativity found in the algorithm.[79] One unique characteristic of these *algo norms* comes from the fact they are embedded in the technology and, as they are structurally conditioned, they cannot be evaded by a digital user.[80]

Another special characteristic comes from the difficult in identifying these norms, not only because they are hidden and opaque, but also because of the theoretical challenge that *algo norms* often reveal themselves in their consequences and we may talk about the existence of a norm only when we realize that there is a normative pattern that needs to be identified, interpreted, and reconstructed.[81] Understanding *algo norms* requires observation of the outcome of the algorithms in real space, in connection with a search for the motives of the relevant actors, their relationships with the algorithms, and the context they create.[82] Importantly, machine learning algorithms adapt according to the data collec-

[77] Hakan Hydén, "Sociology of digital law and artificial intelligence", in Priban, Jiri, *Research Handbook of Sociology of Law*, Cheltenham, Edward Elgar Publishing, 2020, p. 360.

[78] *Ibidem,* p. 361.

[79] *Ibidem,* pp. 361-362.

[80] *Ibidem,* p. 363.

[81] *Ibidem,* p. 364.

[82] *Idem.*

ted and processed in a way that also transforms the normative patterns embedded in them.

In his analysis of future challenges related to artificial intelligence and algorithms, Hydén refers to the same concerns found in the literature related to algocracy: the singularity point in which artificial superintelligence becomes uncontrollable and irreversible, dominating humankind;[83] and societal governance through the introduction of effective regulatory mechanisms of AI[84] According to him, a new and more radical approach to politics, law, and society is necessary.[85]

Our task consists in developing the field of *algorithmic law* to address the challenges posed by *algocracy* and transform these *algo norms* according to our democratic, constitutional and legal standards. The particular character of algorithms demands a special discipline to control their power, architecture, and commands. Nowadays, there is a relevant literature related to the law of robots,[86] but conceptually we should focus on the normative structure of algorithmic law as the core of the discipline.

Our reflection on killer robots, for instance, should be concentrated on the algorithmic programming, instructions, and commands that define the conduct of these military artifacts. Instead of concentrating on the artifacts, we should focus on the algorithmic institutions, that is, the rules of the game and the organizations capable of shaping them. Regarding the prohibition of autonomous weapons, for instance, the focus should be not on the object –the robot, the drone, the weapon, or any other artifact– but rather on the principles and rules embedded in the *algo norms*

[83] *Ibidem,* pp. 365-366.

[84] *Ibidem,* pp. 366-367.

[85] *Ibidem,* p. 367.

[86] See, for instance, Turner, Jacob, *Robot Rules: Regulating Artificial Intelligence*, Cham, Palgrave Macmillan, 2019; Lin, Patrick, Jenkins Ryan & Abney, Keith, (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Oxford, Oxford University Press, 2017.

DOI: http:// dx.doi.org/10.22201/fder.24488933e.2021.279-1.78811

and the algorithmic laws that may change them. These algorithms are behind the radical transformation in the character of this new warfare.[87] Particularly machine learning algorithms are triggering new debates on the borderlands of law and ethics.[88] To be true, the fact that the artifact is often a killer drone is also important,[89] but the relevant moral issues remain the terms of accountability, legitimacy, and fairness of their operations.[90]

For instance, even if the current technology does not support fully autonomous killings because existing systems are not yet capable of distinguishing between a combatant and a non-combatant at a war zone, technological advances made it only a matter of time to develop fully autonomous *human-out-of-the-loop weapons*.[91] Moreover, unless there are clear laws banning the possibility of *human-out-of-the-loop weapons*, there will be inevitable pressure for Autonomous Weapons Systems (AWS) to operate without human supervision, especially because the tempo of the battlefield and the high costs associated with keeping a human 'in the loop' in battles against autonomous opponents.[92]

On the other hand, some experts actually consider the precision, proportionality, and compliance of these weapons to reduce risks of unjustified deaths, facilitate involvement in humanitarian

[87] Elliot, Anthony, *Automated Mobilities: From Weaponized Drones to Killer Bots. Journal of Sociology*, vol.55, num. 1, 2019, p. 28-29.

[88] O'Connell, Mary Ellen, *21st Century Arm Control Challenges: Drones, Cyber Weapons, Killer Robots*, and WMDs, Washington University Global Studies Law Review, vol.13, 2015, p. 526.

[89] Mayer, Michael, *The New Killer Drones: Understanding the Strategic Implications of Next-Generation Unmanned Combat Aerial Vehicles*, International Affairs, vol. 91, 2015, p. 768.

[90] Whetham, David, "Killer Drones: The Moral Ups and Downs", *RUSI Journal*, vol. 158, n. 3, 2013.

[91] Idem, p. 23.

[92] Sparrow, Robert, "Killer Robots", *Journal of Applied Philosophy*, vol.24, n. 1, 2007, .pp. 68-69

intervention and responses to perceived aggression without the need for a full-scale war.[93] In this case, however, the ability of these new technologies to comply with the law of war (*Jus in Bello*) is decisive.[94] If these autonomous weapons meet special technical standards of reliability and may comply with International Humanitarian Law and International Human Rights Law, some would policy experts would support their deployment.[95]

Importantly, once we consider the democratic legitimacy of deployment of these autonomous weapons, we should also acknowledge that public opinion may eventually support killer robots based on the political justifications that are presented to the press by the military.[96] In the end, politics has a role to play in making the fundamental political decisions to establish the algorithmic institutions that will analyze autonomous weapons and define the algorithmic law that will govern the regulation of killer drones and set the rules of the game for development of these weapons.[97] The relevant point is that law and government can channel the

[93] Cfr. Statman, Daniel, "Drones and Robots: On the Changing Practice of Warfare", Lazar Seth and Helen Frowe (eds.), *The Oxford Handbook of Ethics and War,* Oxford, Oxford University Press, 2015.

[94] *Idem.*

[95] Muller, Vincent, "Autonomous Killer Robots Are Probably Good News", in Di Nucci, Ezio and Filippo Santoni di Sio (eds.), *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*, London, Ashgate, 2016.

[96] Michael C. Horowitz, *Public Opinion and the Politics of the Killer Robots Debate*, Research and Politics, 2016*;* Ramanazi, Vaheed, *Killer Drones, Legal Ethics, and the Inconvenient Referent*. Lateral, vol. 7, n. 2, 2018.

[97] Franke, Ulrike Esther, "Drones, Drone Strikes, and U.S. Policy: The Politics of Unmaned Aerial Vehicles", *Parameters,* vol. 44, n. 1, 2014; *A World of Killer Apps, Nature*, vol. 477, 2011; Sandvik, Kristin Bergtora, "The Political and Moral Economies of Dual Technology Transfers: Arming Police Drones", in A. Zavrsnik (ed.), *Drones and Unmanned Aerial Systems*, Cham, Springer, 2016.

development and use of killer robots,[98] eventually even setting a global prohibition of autonomous weapons aligned with the well-known Asimov's three laws of robotics.[99]

Likewise, law and government should also play a relevant role in controlling the development of artificial superintelligence and preventing that machine learning algorithms take over government and control humankind. Revisiting the theme of *algocracy* in a more recent study, Danaher refers to it as the ubiquitous use of computer-coded algorithms to control our world and considers it as a synonym for 'algorithmic governance', 'algorithmic regulation', or 'algorithmic governmentality'.[100] In contrast to his original article with a focus on the threats, Danaher argue this time that *algocracy* may also positively impact our freedom and be emancipatory.[101]

Revisiting the conceptualization of *algocracy*, Danaher reaffirms that it may mean a governance system or an expression analogous to 'democracy' and indicating the idea of a 'rule by algorithm'.[102] According to him, 'algocracy' captures the authority of algorithmically coded architectures in contemporary life.[103] This governance system may reduce our freedom of action when it induces an automatized response like following the instructions to tick a box in the online contract environment.[104] Algorithms may also function like subtle forms of manipulation or

[98] Crootof, Rebecca, "The Killer Robots are Here: Legal and Policy Implications" *Cardozo Law Review*, vol. 36, 2015, p. 1837.

[99] Isaac Asimov, *I, Robot*, Spectra, 2004.

[100] Danaher, John, "Freedom in an age of Algocracy," in Vallor Shannon (ed.), *Oxford Handbook on the Philosophy of Technology*, Oxford, Oxford University Press, forthcoming.

[101] *Idem.*

[102] *Idem.*

[103] *Idem.*

[104] *Idem.*

'hypernudges'.[105] Additionally, algorithms may enable domination on a mass scale.[106]

Danaher also explains that some mechanisms may support algocratic systems in promoting freedom: (1) *choice filtration* – identification and selection of options that might be conducive to your goals, ordering salient patterns in the chaos of data and bringing them to the attention of the user; (2) *cognitive slack*– providing escape routes from cognitive tunnels that lead your choices, because of limitations related to behavioral, ideological or material scarcity that may reduce the ability to see the broad picture, focus on tasks, solve problems, exercise control, and so on.[107] Importantly, however, the complexity of these algorithms mean that they have to be assessed and determined on a case-by-case basis in a properly contextualized manner.[108]

Additionally to algorithmic governance, the risks posed by algocracy also include the threat of government by artificial superintelligence. As Nick Bostrom explains it, the control problem consists of an unprecedented challenge of solving a principal-agent problem in which the human project manages to exercise political control over the superintelligence system.[109] In this context, new techniques are needed to establish proper capability control methods: *Boxing Methods* may segregate physically the system to a box or restrict circulation of information outside the box, so that artificial superintelligence may not have access to physical manipulators or to communications network outside of the box; *Incentive Methods* provide instrumental reasons for an agent to act in ways that promote the principal's interests within an environment; *Stunting* limits the system's intellectual faculties

[105] *Idem.*

[106] *Idem.*

[107] *Idem.*

[108] *Idem.*

[109] Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014, p. 157.

or its access to information; *Tripwires* perform diagnostic tests on the system and effects a shut down if signs of dangerous activities are detected.[110]

Nick Bostrom also considers that motivation selection methods may be used to prevent undesirable outcomes by shaping what the superintelligence wants to do.[111] *Direct specification* consists in defining directly a specific set of rules that would cause artificial superintelligence to act safely and beneficially to mankind.[112] Importantly, however, Nick Bostrom reminds us that the formulation of a highly complex set of detailed rules that applies across a highly diverse sets of circumstances and provides rights responses to all questions is humanly impossible – as evidenced by the legal system with its gaps, revisions, and applications of general common sense.[113]

An alternative would be the method of *indirect normativity* through the definition of a process for deriving standards and the establishment of a system that will pursue this process and implement the standard effectively.[114] Nick Bostrom advocates this method as potential enabler to delegate to the superintelligence cognitive work and reason needed to selected the value to be realized – as part of a strategy of epistemic deference to the superintelligence.[115]

His final recommendation encourages commitment to a common good principle – "superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals".[116] Initially, this principle could come from a voluntary moral commitment individuals and organizations

[110] *Ibidem,* p. 157-169.

[111] Idem, p. 169.

[112] *Idem.*

[113] *Ibidem,* p. 170-171.

[114] *Ibidem,* p. 173.

[115] *Ibidem,* p. 258.

[116] *Ibidem,* p. 312.

within the AI community, but that should be enacted into law and treaty as a sharpened set of specific verifiable requirements.[117]

Our inevitable conclusion should be that algorithmic law emerges as a necessary response to *algocracy*. Governance by algorithms requires the application of algorithmic law to incorporate democracy, rule of law, and fundamental rights into the *algo norms*. Otherwise, our contemporary digital societies will be left with commands, instructions, and programs that harm our polities, institutions, interests, and values. Likewise, the threat of government by algorithms posed by the emergence of an artificial superintelligence that dominates humankind also requires the development of a new algorithmic law that establishes checks and balances and controls the technological system.

Interestingly, in his TED Talk titled "What happens when our computers get smarter than we are?, Nick Bostrom teased his audience with a photograph from a *Terminator* robot. Perhaps there is a more fundamental connection between breeding killer robots and being decimated by artificial superintelligence in the sense that training machine learning algorithms to kill people may provide leverage for the machines to extrapolate this conduct to all humanity once they develop general artificial intelligence and become superintelligent. This is definitely an outcome that we should avoid and algotithmic law should help us designing indirect normativity that foster ethical values and nurture human standards into the superintelligence actors.

IV. Concluding Remarks

'Hasta la vista, baby' means essentially 'see you' or a goodbye. Concluding remarks are an academic form of expressing goodbye too. This article also contains reflections on a potential goodbye to our contemporary form of existence, as we witness the rise of

[117] *Ibidem*, p. 313.

killer robots and artificial superintelligence. Our contemporary digital societies have developed machine learning algorithms with embedded *algo norms* that have the capacity to govern our lives. However, replacing law with mathematical formulas[118] and substituting judgments by algorithmic decision-making may be highly problematic.[119] In this context, law and government should retain their role of providing channels for fundamental decisions, checks and balances for political control, and guarantees that our ethical standards and rules are preserved.

States have the responsibility of developing algorithmic institutions –organizations and rules of the game– that require algorithms to incorporate responsibility, transparency, auditability, incorruptibility, and predictability. Likewise, machine learning algorithms should not learn to dominate humans. In the end, the best strategy to deal with the risks of *algocracy*, killer robots, and artificial superintelligence is to continue to govern ourselves through democracy, rule of law, and protection of fundamental rights. The development of a new algorithmic law could support us towards the mission of maintaining our self-governance and self-government.

[118] Borges Fortes, Pedro Rubim, "How Legal Indicators influence a justice system and judicial behavior: The Brazilian National Council of Justice and 'Justice in Numbers'", *The Journal of Legal Pluralism and Unofficial Law*, vol. 47, n. 1, 2015, pp. 39-55.

[119] Borges Fortes, Pedro Rubim, "Paths to Digital Justice: Judicial Robots, Algorithmic Decision-Making, and Due Process" *Asian Journal of Law and Society*, 1-17, 2020.