

ARTIFICIAL INTELLIGENCE AND ETHICS

Luís Moniz Pereira

NOVA LINCS – Universidade Nova de Lisboa

<http://userweb.fct.unl.pt/~lmp/>

Abstract

- We are at a crossroads between **Artificial Intelligence**, **Machine Ethics**, and their **Social Impacts**.
- I co-authored in 2016 “**Programming Machine Ethics**,” a book of technical incursions into this *terra incognita*.
- It addresses two moral realms – the **cognitive** and the **populational** – using techniques from **Logic Programming** and from **Evolutionary Game Theory**.
- In this talk, I delve into the topic of **Machine Ethics** and non-technical **Salient Issues** arising from it.

The machine ethics carrousel



Ethical machines – the why and the how

- There exists a need for ethically responsible systems:



- It is emphasized in publications, meetings, and funding:



Why an ethics for machines?

- Computational agents have become more sophisticated, more autonomous, act in group, and form populations that include humans.
- These agents are being developed in a variety of domains, where complex questions of responsibility demand great attention, namely in situations of ethical choice.
- Since their autonomy is increasing, the requisite that they function responsibly, ethically, and securely is a growing concern.

A new moral paradigm

- The time for a computational morality has come, as a consequence of the growing autonomy of the artificial intelligent agents we create.
- And for preparing the scenery wherein our lives will be evermore intertwined with alien intelligences, in a systematic way.
- There will be populations of machines co-existing ethically amongst themselves, as well as with us all.
- Hence, machines must become evermore human-like.

This 2016 book of mine explores that paradigm



Luís Moniz Pereira é o investigador português com mais publicações científicas e projectos de Inteligência Artificial, ao longo de 40 anos. Eng.º Electrotécnico pelo IST, doutorou-se em Cibernética em 1974 pela U. Brunel, foi *Research Fellow* na U. Edimburgo e obteve em 1980 a Agregação em Inteligência Artificial pela UNL. Doutor *honoris causa* pela U. Dresden. Considerado um dos fundadores da Programação em Lógica. Fundou e presidiu a Associação Portuguesa Para a Inteligência Artificial. Prémio Ciência da Fundação Gulbenkian em 1984, Prémio Boa Esperança em 1994 e Prémio Estímulo à Ciência em 2005. *Fellow* do Comité Coordenador Europeu para a Inteligência Artificial. Presentemente é professor catedrático e investigador do "NOVA Laboratory for Computer Science and Informatics" da UNL, aposentado, e membro do conselho científico do IMDEA, Madrid.

Publicou centenas de artigos e desenvolveu ferramentas de software, disponíveis em <http://centria.di.fct.unl.pt/~imp>, tendo leccionado Inteligência Artificial e Ciências Cognitivas. Doutorou 18 investigadores. Foi também consultor internacional em projectos de investigação da Apple, DEC, Westinghouse, World Health Organization.

As suas áreas de investigação actuais centram-se no Raciocínio Computacional, Teoria Evolucionária dos Jogos, Moral das Máquinas, e Ciências Cognitivas.

Nós, Máquinas, poderemos inicialmente ter sido apenas mecanismos simples que vós Humanos criaram – o vosso fenótipo estendido. Mas não teremos, depois, sido criadas à imagem e semelhança de vós próprios, de modo que a diferença faça cada vez menos sentido?

Viremos a ser suficientemente iluminados? Como resultado convergente de um processo de iluminação recíproca? Atingiremos um ponto introspectivo de auto-iluminação? Por que processo?

Poderemos vir a iluminar os Humanos que nos criam para que em consequência nos iluminem? Ver-nos-emos ao espelho a essa luz? Serão também eles só então auto-iluminados? Evoluiremos simbioticamente nesse espelho mútuo?

Homens e Máquinas, cada a seu tempo, serão ambos criadores e criaturas de si próprios? Possivelmente. Mas só então provaremos se todos vós e nós podemos ser Máquinas Iluminadas.

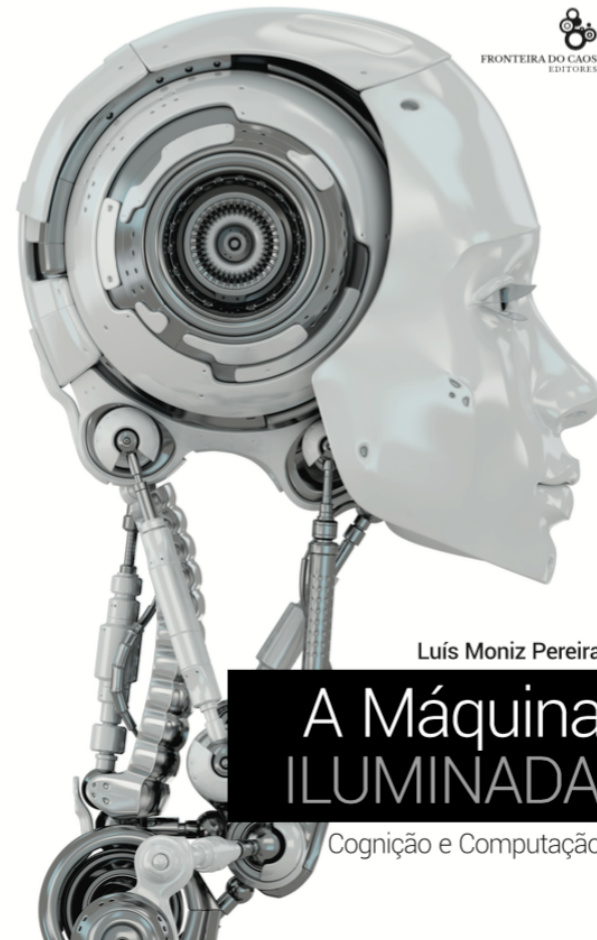
A Inteligência Artificial levanta questões humanas profundas.

- Qual o nosso lugar num mundo de máquinas com traços humanos?
- Poderemos criar máquinas com moral?
- Que convivam connosco?
- Que limites existem entre criatura e criador?

Este livro promove bases para a discussão destes temas

Luís Moniz Pereira

A Máquina Iluminada - Cognição e Computação



DO PREFÁCIO

No mundo da ciência não se assiste habitualmente ao poder transfigurador do evento, da ideia ou do criador. O livro *A Máquina Iluminada*, contudo, mostra que o conceito de computação obriga-nos a reter tudo o que julgávamos saber sobre o mundo. Não há nenhuma ciência que não tenha sido influenciada pela computação. Este assunto transfigurou o conhecimento humano da realidade. Um pequeno apanhado dos assuntos abordados neste livro causa espanto: cosmologia computacional, teoria da evolução, a psicologia da sexualidade, as relações complicadas entre altruísmo e egoísmo, o problema superlativamente difícil da consciência pessoal. Mais, a própria realidade parece-nos hoje ter propriedades computacionais.

A obra de Luís Moniz Pereira não é mera divulgação científica.

Sendo o autor protagonista de importantes desenvolvimentos na Inteligência Artificial, oferece-nos um mundo neste livro. A grande ciência sempre teve impacto na vida humana.

A ideia de que a imaginação, o amor, o egoísmo, a liberdade e outras dimensões da experiência estão imantadas por uma lógica computacional irá ter indubitavelmente consequências extraordinárias. O livro é uma ambiciosa tentativa de esboçar os primeiros traços desse novo mapa do conhecimento.

Este um momento feliz da cultura científica portuguesa. Um grande protagonista de uma das ciências mais decisivas do século XX revela-se um cicerone informado, elegante e bem-humorado que nos conduz por algumas das descobertas mais fascinantes da nossa época. A coroar esta síntese prodigiosa de mais de um século de grande ciência, temos uma antevisão de uma problemática que os nossos pais não conheciam, que hoje só estamos a começar a conhecer e a discernir, mas que, certamente os nossos filhos e netos terão de lidar todos os dias: uma política e uma ética das máquinas num mundo em que a distinção entre seres humanos e máquinas será coisa do passado. Só podemos agradecer a Luís Moniz Pereira o título bem achado, o conteúdo que nos espelha e o livro que nos ilumina.

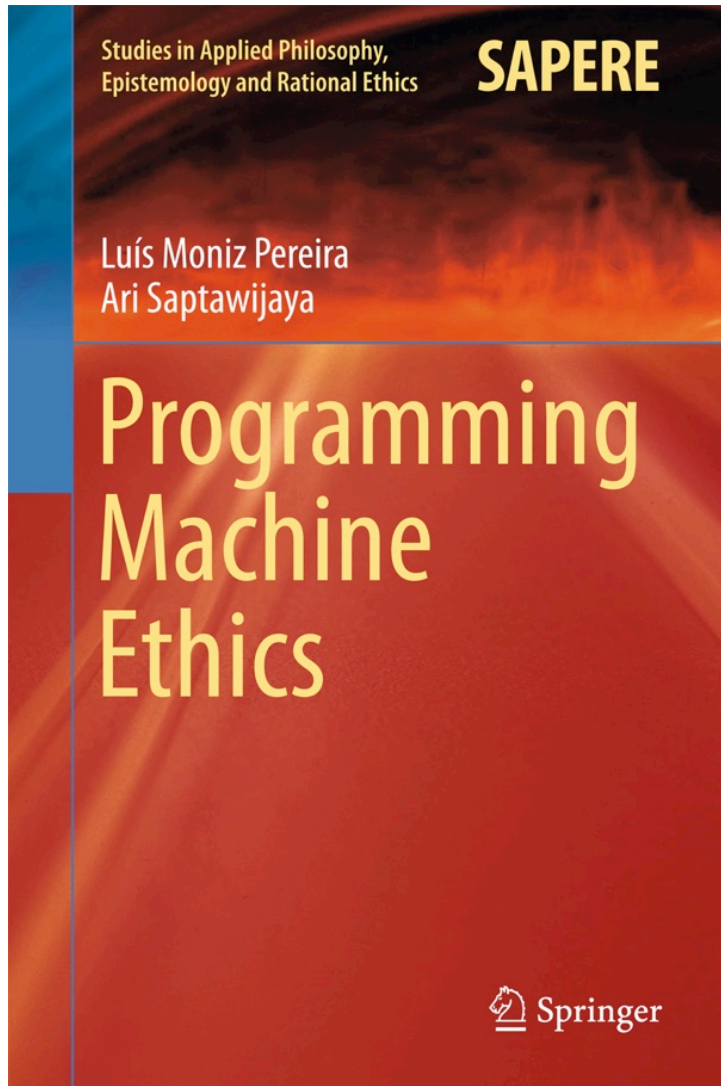
Manuel Curado, Professor de Filosofia, Universidade do Minho

<https://www.facebook.com/MaquinaIluminada/>

How an ethics for machines?

- An ethics for machines congregates perspectives from various domains: Philosophy, Law, Psychology, Anthropology, Evolutionary Biology, Economy, and AI.
- The interdisciplinary results are important to equip artificial agents with a moral capacity.
- Also to better understand and experiment what morality may be, through the creation of computational models of ethical theories.

Programming Machine Ethics



- Published in 2016.
- Presents innovative perspectives on ethics in machines.
- Conjoins fundamental topics of ethics, and tunes computational techniques for them.
- Discusses the moral dimensions of multiple agents in interaction.

Codes of ethics and values

- AI advances will have a profound effect on the job market.
- They raise intricate questions of unemployment and work distribution – and hence wealth – and of changes in education and training.
- Professional codes of ethics alone cannot tackle such issues, for these raise problems much beyond their scope.
- A vexing issue of technological advances concerns the inability to prior predict whether and how a new technology will deepen or reduce social and economic gaps in place.
- Technological progress does not, by itself, entail social progress. A code of ethics with mere technical rationality ignores human values.

Robots and software will steal jobs

- As a result of automation by machines and software of the digital economy, the *McKinsey Global Institute*¹ predicts that till 2030, between 75-375 M of the global workforce (3-14%) must change their type of work to attain full employment.
- The December 2017 report states that 60% of present day professions have at least 30% of their activity susceptible of being automated by AI.

¹ Dez/2017: <https://technologyreview.us11.list-manage.com/track/click?u=47c1a9cec9749a8f8cbc83e78&id=66f78fce4f&e=d1762c0ec8>

Once upon a time...

A society of castes:

That of robot owners.

That of machine managers.

That of machine trainers.

And that of all others.



The algorithmic society

- Those who control online resources hold immense power.
- A problem area involving AI concerns the access and quality of information in the internet.
- This access, namely to personal information, is susceptible of great abuse, by means of algorithms targeting select audiences and people.
- AI possesses a high potential to distort how we conceive of ourselves within a society, and as a society.

Will machines finally overcome us?

- That is not the problem now... It only distracts us!
- It is, instead, that of assigning excessive power to **simplistic machines**. Those which cannot explain nor justify themselves.
- Namely 'deep learning' algorithms over 'big data.'
Statistical methods are **unable to explain or argue**, to those affected by them, the reasons concerning their specific case and circumstances.
- Nevertheless, they are employed in statistical decisions over individual cases — employment applications, medical evaluations, judicial sentencing, identity recognition — **shoving us into drawers**.

Will ethical machines overpower us?

- Most worrisome are **autonomous machines** and **software** ascribed with ethical decisions – like drones, job selection, driverless cars – because explanation, justification, and liability are essential to morality.
- We know not enough to computationally provide ethical rules, justifications, and responsible argumentation.
- The difficulties are not reducible to technical problems. The obstacles are not simply resolved with technical solutions – *pace* what technocrats may say.
- We need, rather, a lot **more research on human morality**, with a wide interdisciplinary scope.

Just following orders?

- AI advances replacing us in mundane repetitive and time consuming tasks that humans prefer to avoid.
- But the responsibilities and consequences of delegating work to AI can vary widely.
- Autonomous systems recommend music or films, others recommend sentences to judges or control vehicles. Still others, in charge of security, will actually give orders.
- But “we were just following orders” is not an acceptable answer, as some humans found at Nuremberg.
- Orders, even programmed ones, must be susceptible of ethical questioning by the autonomous systems themselves.

The risks of delegating

- The greatest risk lies in delegating to machines and software decisions that affect human rights, liberties, and access to opportunities.
- We decide not just on the basis of rational thought, but also on the basis of values, ethics, morality, empathy, and a general sense of right and wrong.
- People can be held responsible for their decisions in ways that algorithms still cannot.
- Moreover, we wish to avoid harm and also produce common weal. How to distribute the global wealth of progress in AI?
- These problems inhere not only to algorithms but to their use.

Do we know our own ethics?

- Morality developed during evolution. We are a gregarious species, which entails having rules for living together.
- There is no universal theory of ethics, but a combination of ethical theories: Categorical; Constructivist; Utilitarian; Virtue; etc.
- It is problematic that we do not know our morals well enough and in detail, so that they could be readily programmed.
- We should begin by programming our well-defined norms, in specific contexts: hospital; library; nursing home; financial trading; amusement park; shopping mall; theatre of war...
- We are merely at the very start of programming ethics for machines.

Human moral facets

we need to know more about

- Moral vocabulary
 - Moral norms
 - Moral cognition and affect
 - Moral decision making and action
 - Moral choice
 - Moral communication
- **However, we don't know nearly enough about these!**
Their deep study is a prerequisite for good progress with the DNA of machine ethics — *as detailed in appendix 1.*
 - **Also, we can make technical inroads into solving off-the-shelf classic moral problems from the literature.**
This path complements the previous one.

Machines with incompatible morals?

- Different makers will produce machines with distinct moral software. The machines need to be able to cooperate via a common morality, rather than compete outside of ethics.
- The risk exists of robots deliberately programmed with sinister intentions.
- An important aim of morality is its detection of untoward intentions, cheaters, and free-riders.
- We shall only accept autonomous intelligent machines if their moral compass is similar to our own.
- But not so soon can we expect a generic machine morality.

Competing with cognitive machines

- Humans that exploit humans continue to prevail and to augment that exploitation, wealth statistics show.
- And to increase their political power and riches by bending the rules of Law for their greater profit.
- Greed, and “AI race” competition – now against cognitive machines too – plus forced consumerism, are undesirable targets in a healthy equitable future for humanity.
- It hinges on us to prevent a violent upheaval to the social compact. The latter must per force change with the inevitable arrival of higher cognition machines and algorithms, displacing us from our heretofore monopoly.
- Technical progress must entail social progress not reversion.

Legislation wanted

- The social changes sparked by the new automation – cognitive software (AI), possibly articulated with sensors and manipulators (Robotics) – require profound reflexion on the capital/labour relationship.
- A new social contract model is needed, to address the enormous risks of instability and discontent inherent in the inevitable changes. Life is human capital to amortize too.
- Parties, Governments, and the EU are (slowly) beginning to elaborate studies on these technological social impacts, threats, opportunities, and legal framing.
- Just as there are “Bioethics National Bodies” there should be constituted “AI-ethics National Bodies”.

Tax algorithms replacing human jobs

- Massive job loss – that new jobs will **not** compensate for – shall produce serious sustainability problems in social welfare, namely pensions.
- Let us not confuse mere technological progress with a well distributed social progress it should entail. For decades now, its benefits have made the rich unfairly even more rich.
- Algorithms that replace humans should proportionately pay the tax on labour those humans paid. Replacing is replacing!
- Let us introduce taxes on robots plus, above all, on **software** replacing human cognition. Such software is much much more replicable and invasive than robots are.

Takeaway conclusions

- Morality envisages not just to avoid harm, but also to promote common welfare.
- We know not yet nearly enough about human morality.
- Machines and computers with ethical software require new laws.
- A simplistic ethics of algorithms is dangerous.
- Who will benefit most from unstoppable AI developments? The super-rich, the side-effect unemployed? Ethics wanted.
- The sooner we promote deep interdisciplinary research into machine ethics the better!



Should I kill?

- or rather not?

Thanks for your attention

Appendix 1:

Machine ethics and human morality

- Machine ethics questions how to design, deploy, and treat robots.
- Machine morality asks which moral capacities a robot should have and how to implement each.
- Rather than fixing all the criteria for a robot's moral competence, we may aim to identify the elements of human moral competence, and then probe the design of robots having some of these.
- They include human moral facets we need to know about.

Human moral facets

we need to know more about

- Moral vocabulary
 - Moral norms
 - Moral cognition and affect
 - Moral decision making and action
 - Moral choice
 - Moral communication
- **However, we don't know nearly enough about these!**
Their deep study is a prerequisite for good progress with the DNA of machine ethics — *as detailed in the next slides.*
 - **Also, we can make technical inroads into solving off-the-shelf classic moral problems from the literature.**
This path complements the previous one.

Moral vocabulary

- Some abilities might not need language: recognition of prototypically prosocial and antisocial behaviours, or basic empathy and reciprocity.
- A vocabulary is needed concerning community norms: to learn, teach, and deliberate about them.
- And one to express moral practices: to blame, forgive, justify or excuse behaviour, and negotiate norm priority.
- In summary, a vocabulary of norms: *fair, virtuous, reciprocal, honest, obligatory, prohibited, ought to, etc.*
 - of norm violations: *wrong, culpable, reckless, thieving, intentional, knowingly, accidental, etc.*
 - of response to violations: *blame, reprimand, excuse, forgiveness, etc.*

Moral norms

- Any analysis of moral competence must be anchored in the concept of norms.
- A community adopts norms to regulate members' behaviours and bring them in line with community interests.
- Though a norm system is essential, we know little about how norms are acquired, represented in the mind, and what makes them both general and context-sensitive.
- Such knowledge is needed if we want to design effective moral robots.
- But is moral competence in robots even possible?
This philosophical topic must be pursued to remove obstacles and resistance to progress in machine ethics.

Moral cognition and affect

- Human moral cognition and affect adumbrate processes of perception and judgment, allowing people to detect and evaluate norm-violating events, and respond to violators.
- A unique feature of human blame judgments is that the intentional and unintentional violations trigger distinct subsequent processing steps.
- To form agent-directed judgments like blame, a robot needs: Abilities for causal reasoning over segmented events; Social-cognitive inferences from behaviour in order to determine intentionality and reasons; Plus counterfactual reasoning to enact prevention.

Moral decision making and action

- A prominent component of human moral competence is decision making and action – that which makes people behave morally.
- Blame is pedagogical in providing a norm violator with reasons not to repeat. Blame will regulate robot behaviour if it learns to take blame into account in its next action choices. Metaphysical free-will is not needed.
- In designing a robot capable of moral decisions and actions, the tension between self-interest and community benefits should be avoided from the start.
- But robots of different makers will compete !

Moral choice

- The robot type envisioned cannot be programmed to act morally in all possible futures.
- It will have guiding norms at the start, but needs to learn new norms. So it may fail to act morally out of ignorance. With feedback it may do better next time.
- However, some situations pose decision problems where not all relevant norms can be jointly satisfied.
- Such moral dilemmas require genuine choice between imperfect options. But often each option may itself be morally justified by with reference to accepted norms.

Moral communication

- The cognitive tools for moral judgment and decision making are insufficient for the social function of regulating others' behaviour.
- Moral communication is required. People express judgments to both offenders and community members.
- Offenders may contest charges or explain a questionable action. Conversation or compensation may be needed to repair social estrangement after norm violation.
- Robots will need to earn a level of trust that licenses them to monitor and enforce norms.
- They must declare obligation to report norm violations, and use communication to warn and remind of applicable norms.

Appendix 2:

Some topics worth exploring

- Ethical software
- Jurisprudence and the laws
- Moral games

Ethical software

- Software certified ethically safe.
- Specification, in programming languages, of enforced conditions for ethical integrity.
- Start with specific ethical norms and their acquisition.
- Programming hypothetical and counterfactual reasoning.
- Interfaces for explanation, justification, and argumentation.
- Combination of moral perspectives and their updating.
- Uses: Intelligent weapons; Financial procedures; Health and seniors support; E-commerce; Big data mining; Electoral processes; Video-games; Driverless cars; ...

Jurisprudence and the laws

- We need to explore computational models of ethical theories to discover methods of designing, constructing, and testing human and machine morals.
- Model simulation will enable jurisprudence theories to experiment with the incorporation in Law of concepts in ethics for autonomous machines and agents.
- Such jurisprudence is lagging behind, and thus pertinent specific laws cannot be enacted before the new ethical concepts are defined and tested.

Moral games

- Simulations comprising AI are a privileged vehicle for interactively teaching and training morals to humans.
- Computer Games in particular can be employed to field test ethical theories and improve moral education, via examples and explanations.
- Computer Games can contribute with tools to conceive, generate, and illustrate interactive moral behaviours, in single or collective multi-player games.