# FROM MACHINE MORALS TO THE MACHINERY OF MORALITY

**Luís Moniz Pereira**

NOVA LINCS – Universidade Nova de Lisboa

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

Lions Clube, Guimarães, Portugal

16 April 2021

# Abstract

- We are at a crossroads between **Artificial Intelligence**, **Machine Ethics**, and their **Social Impacts**.

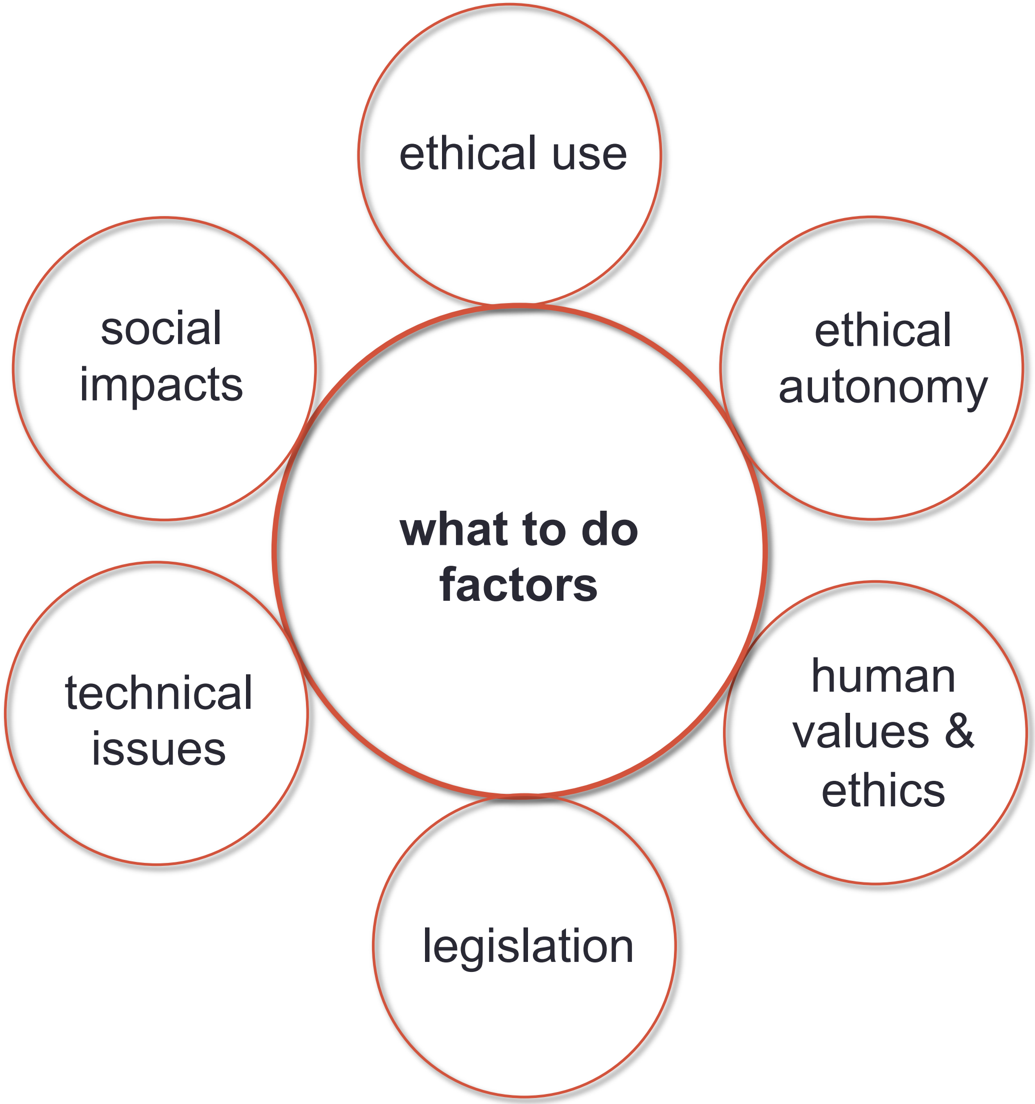
- I co-authored in 2016 “**Programming Machine Ethics,**” a book of technical incursions into this *terra incognita*.

- It addresses two moral realms – the cognitive and the populational – using techniques from **Logic Programming** and from **Evolutionary Game Theory.**

- This talk delves into machine ethics and the non-technical issues arising from it. Covered in a 2020 book I co-authored:

**Machine Ethics** (English) & **Máquinas Éticas** (Portuguese)

# The machine ethics carousel



what to do factors

ethical use

ethical autonomy

human values & ethics

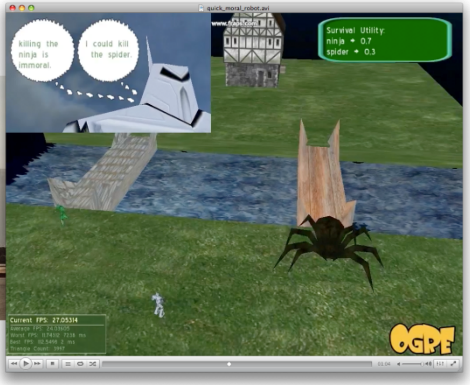
legislation

technical issues

social impacts

# Ethical machines – the why and the how

- There exists a need for ethically responsible systems:



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

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 523 > Issue 7558 > News Feature > Article

NATURE | NEWS FEATURE

Machine ethics: The robot's dilemma

Working out how to build ethical robots is one of the thorniest challenges in artificial intelligence.

Boer Deng

01 July 2015

US Navy funds morality lessons for robots

14 MAY 14 / by CHRIS HIGGINS

As we all learned from the 1986 film *War Games*, machines have the upperhand in warfare when it comes to making logical decisions (such as, the only winning move in nuclear war is not to play). But now it seems the US Navy is not content with that party trick, as it is working on teaching artificial intelligence how to make moral and ethical decisions, too.

A multidisciplinary team at Tufts and Brown Universities, along

# 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. Engº 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/~lmp, tendo leccionado Inteligência Artificial e Ciências Cognitivas. Doutorou 18 investigadores. Foi também consultor internacional em projetos 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 iluminadas? 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

A Máquina Iluminada - Cognição e Computação

Luís Moniz Pereira

FRONTEIRA DO CAOS
EDITORES

FRONTEIRA DO CAOS
EDITORES

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 reler 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 irmanadas 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.

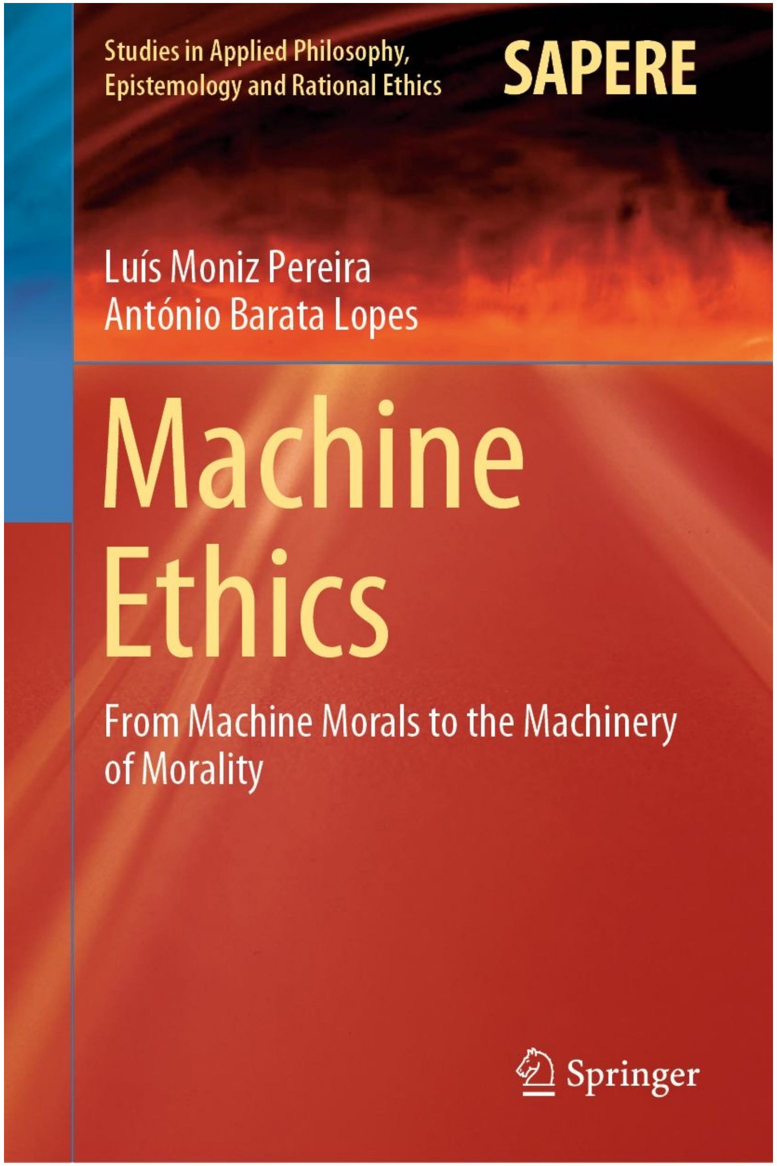
Manuel Curado, Professor de Filosofia, Universidade do Minho

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.

# My recent 2020 book deals with Machine Ethics



Studies in Applied Philosophy, Epistemology and Rational Ethics

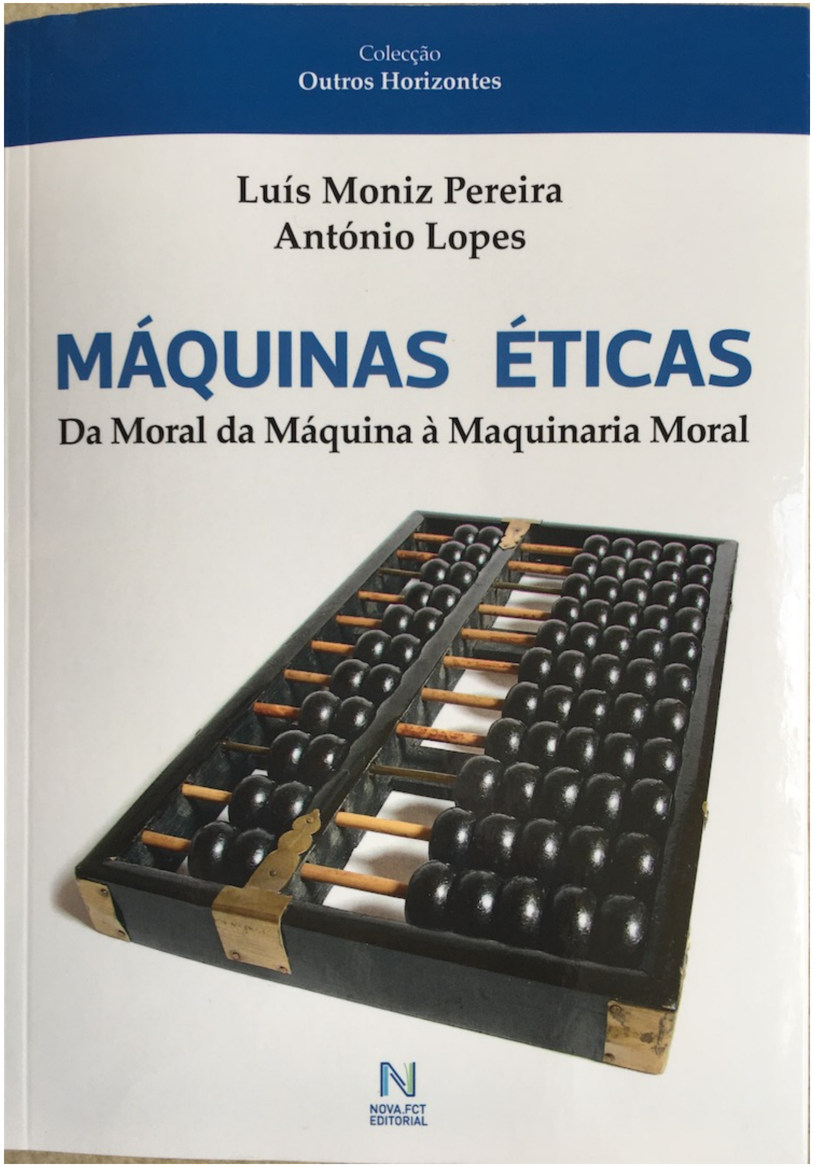
SAPERE

Luís Moniz Pereira
António Barata Lopes

Machine Ethics

From Machine Morals to the Machinery of Morality

Springer



Colecção
Outros Horizontes

Luís Moniz Pereira
António Lopes

MÁQUINAS ÉTICAS

Da Moral da Máquina à Maquinaria Moral

NOVA.FCT
EDITORIAL

# Two Realms of Machine Ethics

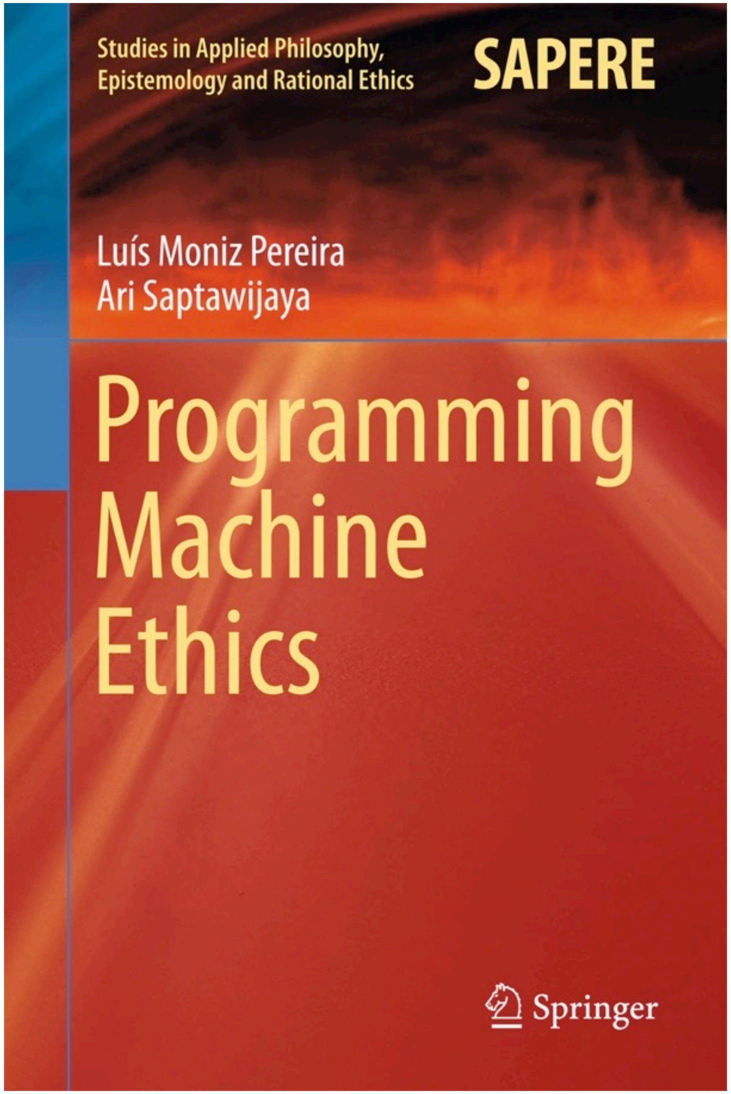
- Our research contemplates two distinct realms of machine ethics − the individual and the collective − identifying bridges between them.

- In the individual realm, we focus on Logic Programming techniques for modeling moral permissibility; on the dual-process of moral judgments; and on counterfactual reasoning in morality.

- In the collective realm, we focus on the emergence of cooperation in populations — where individuals are equipped with diverse cognitive abilities and behavior strategies — by employing Evolutionary Game Theory techniques.

This more technical book of mine addresses such issues.

# Programming Machine Ethics



Studies in Applied Philosophy, Epistemology and Rational Ethics

SAPERE

Luís Moniz Pereira
Ari Saptawijaya

Programming Machine Ethics

Springer

- 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 and September 2018 reports state that 60% of present day professions have at least 30% of their activity susceptible of being automated by AI.

[1] December 2017: JOBS LOST, JOBS GAINED: WORKFORCE TRANSITIONS IN A TIME OF AUTOMATION
September 2018: Notes from the AI frontier: Modeling the impact of AI on the world economy

# 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.

PRICE $8.99
THE NEW YORKER
OCT. 23, 2017

# 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.

# Beyond Programming Machine Ethics

- Recall we stand at the crossroads of AI, Machine Ethics and their impact on society.
- We must not stop at the prevention of harm, but proceed to the ideological and political topics of promoting general well-being and fairness when using machines and software.
- Overall results are important not just for equipping agents with abilities for moral judgment. But also for helping us understand morality better, via the creation of computational models and testing of theories of ethics.
- Computer models make them well defined, eminently observable in their dynamics, and transformable incrementally in expeditious ways.

# 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 and consent

- 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

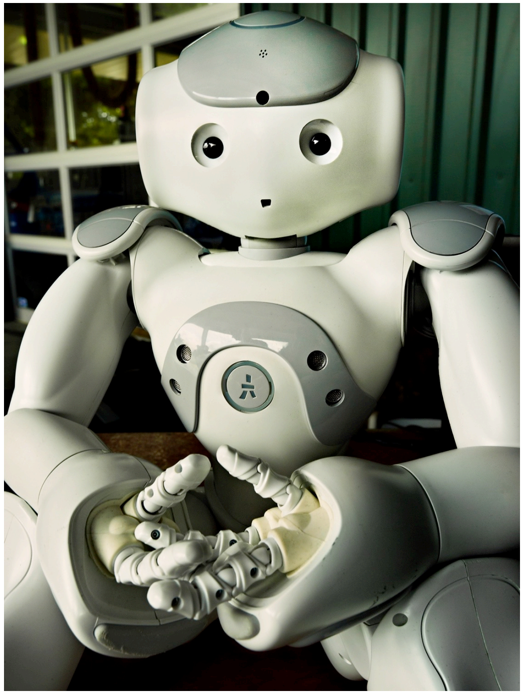
- The social changes sparked by the new automation – cognitive software (AI), possibly articulated with sensors and manipulators (Robotics) – require profound reflexion on the capital/labour relationship.
- A new social contract model is needed, to address the enormous risks of instability and discontent 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!



Many thanks to my co-authors:

| | |
|---|---|
| Ari Saptawijaya | (Indonesia) |
| The Anh Han | (UK) |
| Tom Lenaerts | (Belgium) |
| Francisco C. Santos | (Portugal) |
| Luis Martinez-Vaquero | (Spain) |
| António Barata Lopes | (Portugal) |

L. M. Pereira, Maschinenmoral (in English: Moral Machines), Interview by Nora Saager, P. M. Magazin, pp. 30-35, February 2018

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 and consent

- 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.*
- But 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 and consent

- The cognitive tools for moral judgment and decision making are insufficient for the social function of regulating others' behaviour. Consent is also required.
- Moral communication is every where. 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.

Appendix 3:

# The Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems

# Guilt - 1

- We present models wherein agents may express guilt, to study the role of guilt in promoting pro-social behaviour.

- Analytical and numerical methods from evolutionary game theory (EGT) are employed to find conditions for enhanced cooperation to emerge, within the context of the iterated prisoners dilemma (IPD).

- Guilt is modelled explicitly in guilt prone agents:
  - a counter keeps track of the number of transgressions;
  - a threshold determines if guilt alleviation is performed, by self-punishment and behaviour change to cooperation.

# Guilt - 2

- Alleviation has a subtracting effect on the payoff of a guilty agent.

- If agents resolve their guilt without first considering their co-player's attitude towards guilt alleviation, then cooperation does not emerge:

  Guilt prone agents are dominated by those not experiencing guilt or not acting on it.

- However, cooperation can thrive when a guilt prone agent alleviates her guilt only if guilt alleviation is manifest in a defecting co-player.

# Guilt - 3

- Our analysis provides important insights into the design of multi-agent systems, because inclusion of guilt can improve the agents’ cooperative behaviour, with overall greater benefit as a consequence.

# Guilt - Blame and Punishment

- To prevent blame, there exists a self-punishing guilt mechanism.
- It is associated with *a posteriori* guilt for a harm done, whether or not intended.
- It functions *a priori* too, preventing harm by wishing to avoid guilt.
- The *a posteriori* outward admission of guilt may serve to pre-empt punishment, when harm detection and blame by others becomes foreseeable.

## Appendix 4:

# Counterfactual Thinking

# in Cooperation Dynamics

# Counterfactual Thinking (CT)

- CT is a human cognitive ability studied in a wide variety of domains, namely Psychology, Causality, Justice, Morality, Political History, Literature, Philosophy, Logic, and AI.
- CT captures the process of reasoning about a past event that did not occur, namely what would have happened had the event occurred.
- CT is also used to reason about an event that did occur, concerning what would have followed if it had not.
  Or if another event might have happened in its place.

# Example

An example situation:

- *Lightning hits a forest and a devastating forest fire breaks out. The forest was dry after a long hot summer and many acres were destroyed*.

A counterfactual thought is:

- *If only there had not been lightning, then the forest fire would not have occurred*.

# Evolutionary Game Theory

- Given the wide cognitive empowerment of CT in the human individual, the question arises of how the presence of individuals with CT-enabled strategies affects the evolution of cooperation in a population comprising individuals with diverse strategies.

- The natural locus to examine this issue is Evolutionary Game Theory (EGT), given the amount of extant knowledge concerning different types of games, strategies and techniques for the evolutionary characterization of such populations.

# Adding CT to EGT

- In the context of the social learning model of EGT, individuals revise their strategy by looking for the greater success and actions of others and copying their strategy.

- Yet, contrary to social learning, an agent may instead imagine how an outcome could have turned out if she would have acted differently, and revise her strategy accordingly.

- We propose simple models to study the impact on cooperation of having a fraction of agents resorting to such CT, possibly in a population of social learners.

# The Counterfactual Payoff

- In EGT, a simple CT can be exercised after knowing one's resulting payoff, following from a single playing step with a co-player.
- It employs the counterfactual thought:

  *Had I played differently, would I have obtained a better payoff than the one I did?*

- This payoff information is easily obtained by consulting the game's payoff matrix, while assuming the co-player would keep to the same play;  i.e. other things being equal.
- In the positive case, the CT player will then next adopt the more positive alternative play strategy.

# Adding Theory of Mind to CT in EGT

- A more sophisticated CT would search for a counterfactual play that improves not just one's payoff, but contemplates as well the co-player not becoming worse off, in fear the co-player will react negatively to one's change of strategy.
- More sophisticated still, the new alternative strategy may be searched for by taking into account that the co-player also possesses a CT ability.
- Furthermore, the co-player might too employ a Theory of Mind-like CT, up to some level.
- We examine here only the non-sophisticated case.

# CT and Social Learning (SL)

- CT can be envisaged as a form of strategy update, akin to program debugging and to the best-response rule in game theory, in the sense that:

  *If my actual play move was not conducive to a good payoff, then, after having known the co-player's move, I can imagine how I would have done better had I made a different strategy choice*.

- In EGT, a frequent form of learning is so-called Social Learning (SL). It consists in switching one's strategy from time to time, by imitating the strategy of a more successful individual in the population, rather than using the CT.

# Conclusion

- Counterfactual thinking by individuals in populations has proven worth of study.
- It enables the arising of increased cooperation, even where non or little existed before.

O meu mais recente livro intitula-se

”**Da Moral da Máquina à Maquinaria Moral**"

é da autoria de

**Luís Moniz Pereira** e **António Lopes**
foi publicado pela NOVA.FCT Editorial em 2020

OBJECTIVO

Trata-se de um livro de divulgação científica e índole cultural, intitulado ”*Da Moral da Máquina à Maquinaria Moral*," da autoria de Luís Moniz Pereira (Professor Catedrático aposentado da FCT-UNL, membro do seu centro NOVA-LINCS do Departamento de Informática) e de António Lopes (Mestre e professor de Filosofia no Ensino Secundário público).

Constitui uma obra de divulgação científica e índole cultural, destinada a proporcionar percepções abrangentes sobre um tópico muito actual da Inteligência Artificial (IA). O seu objectivo é o de disponibilizar para um público bastante vasto conteúdos de reflexão vivamente actuais, indicados no seu título, contribuindo para debates muito mais informados sobre o tópico.

O material de base para o realizar é constituído por um conjunto substancial de artigos científicos especializados ultimamente publicados por Luís Moniz Pereira, bem como entrevistas e palestras proferidas por este cientista, praticamente na totalidade em Inglês, e os quais são trabalhados de forma a produzirem um todo coerente, articulado em Português, e adaptado a um público não especializado. A qualidade e pertinência desses materiais justifica a sua publicação em Língua portuguesa.

O formato é o de um diálogo entre um cientista e filósofo – Luís Moniz Pereira – e um filósofo e romancista – António Barata Lopes. Ao recuperar esta forma de exposição clássica – que já vem desde Platão – os autores pretendem dar nota de que todo o conhecimento segue uma lógica de problemas e soluções que, por sua vez, abrem horizontes para novos problemas. Sinaliza também que no conhecimento científico não existem tópicos fechados sobre si próprios; assim sendo, colocar adequadamente uma pergunta já aponta para os modos de soluções possíveis. Por outro lado, tornará muito mais compreensível e mais dinâmica toda a aproximação dos leitores à temática explorada.

A composição da obra articula três dimensões da questão. Em primeiro lugar, uma abordagem ao conceito de inteligência e ao modo como ele evoluiu; em segundo lugar uma abordagem aos tópicos da Economia e sociedade, especulando sobre impactos vários da IA na vida concreta das pessoas. Por fim enfrenta-se a questão especifica da autonomia das máquinas e a necessidade de as dotar com uma moral que lhes permita um criterioso relacionamento entre elas próprias, e delas com os humanos. Pelo meio, endereçam-se questões epistemológicas sobre a formulação da moral em computador, e o estudo por simulações em computador da sua evolução emergente em populações de agentes.

O livro tem a duzentas e dezoito páginas. A primeira parte destina à exploração evolucionária do conceito de inteligência; parte seguinte analisa os variados impactos sociais e económicos evidenciando a necessidade de uma moral social reconfigurada; a parte final é destinada ao tema da moral computacional, sumarizando trabalhos realizados por Luís Moniz Pereira, e explicitando a urgência da investigação e a necessidade de conclusões implementáveis no imediato.

## JUSTIFICAÇÃO

Perante o actual estado da IA, no qual o surgimento de ferramentas de *deep learning* sobre *big data* permite tratar dados numa quantidade e qualidade até agora impensáveis; em que se geram algoritmos cada vez mais capacitados para tomarem decisões autónomas; e é pensável a implementação dessa tecnologia em robôs com várias funções, como máquinas de guerra, automóveis ou aviões, emerge uma questão que é incontornável: Os seres humanos não serão os únicos agentes autónomos, com capacidade para deliberar sobre aspectos que impactam directamente na nossa vida.

Neste contexto, a deliberação autónoma e criteriosa reclama por regras e princípios de natureza moral aplicáveis à relação entre máquinas, à relação entre máquinas e seres humanos e aos impactes resultantes da entrada destas máquinas no mundo do trabalho e na sociedade em geral.

Luís Moniz Pereira tem trabalhado neste domínio ao longo dos últimos 14 anos. Tendo por base um paradigma apoiado nos dados da Psicologia Cognitiva e Moral Evolucionarias, endereçando a moral como um caso da teoria dos jogos evolucionários, produziu um conjunto muito extenso de artigos científicos e outros trabalhos que se encontram maioritariamente em língua inglesa. Estes estudos têm a particularidade de exprimirem uma abordagem científica da moral, simulável em computador, e aplicável ao domínio da moral computacional e social. Ora, urge fazer uma síntese dessa investigação e disponibilizá-la em língua portuguesa para um público leigo nessa matéria. O tema da moral computacional interessa não apenas empresas e instituições públicas, mas também a quem queira exercer uma cidadania consciente e crítica.

O actual estado de desenvolvimento da IA tanto na sua capacidade de elucidação dos processos cognitivos emergentes na evolução, quanto na sua aptidão tecnológica para a concepção e produção de programas informáticos e artefactos inteligentes, constitui-se como o maior desafio intelectual do nosso tempo.

Do ponto de vista do paradigma acerca do que é a evolução e a cognição, as investigações em torno desta área do conhecimento têm evidenciado uma perspectiva muito mais integradora. É possível ver a inteligência como resultado de uma actividade de processamento de informação, e traçar uma linha evolutiva que vai dos genes aos memes, e sua co-evolução. Nestes termos, rupturas tradicionais entre o ser humano e os restantes animais, ou entre cultura e natureza passam a fazer pouco sentido.

Toda a vida é um palco evolucionário, onde a replicação, a reprodução e a recombinação genética têm ensaiado soluções para uma cognição e uma acção cada vez mais aprimoradas e distribuídas. A biologia, dada a sua matriz computacional, instaura sobre a Física uma primeira artificialidade. Assim sendo, o actual estado do conhecimento implica uma redefinição do lugar do ser humano no mundo, lançando desafios a várias áreas do conhecimento. Desde logo a muitas disciplinas da Filosofia, pois problemas como o que é conhecer, o que é o homem, e o que são e como surgiram valores de natureza moral ganham aqui perspectivas até agora impensáveis.

No que diz respeito ao conhecimento, surge a possibilidade de o mesmo ser simulado em computadores, superando desta forma os limites que antes eram impostos por uma especulação que não podia passar da experiência mental, quiçá compartilhada.

No que diz respeito ao questionamento antropológico, a tradicional discussão sobre "O que é o Homem?", mercê do cruzamento entre a IA, a engenharia genética e a nanotecnologia, vê-se agora substituída por uma poderosa e desafiante problemática em torno daquilo que pode vir a considerar-se desejável e possível que seja e irá sendo o Homem.

Do ponto de vista dos critérios de acção, a moral alcandorada nos céus do passado está confrontada com uma nova perspectiva sobre os sistemas morais nascentes, estudados no âmbito da psicologia evolucionária e aprofundados através de modelos testáveis em cenários artificiais, como agora permitido pelos computadores. À medida que a investigação avança, podemos conhecer melhor os processos inerentes à decisão moral, ao ponto de eles poderem ser "ensinados" a máquinas autónomas capacitadas para manifestarem discernimento ético.

No domínio da Economia há toda uma problemática associada ao impacte no trabalho e à dignidade que lhe é inerente, bem assim como à produção e distribuição da riqueza; ou seja, toda uma reconfiguração das relações económicas que resultará não apenas da automação de actividades rotineiras, mas fundamentalmente da entrada em cena de robôs e software que poderão substituir médicos, professores, ou assistentes em lares de terceira-idade (para darmos nota de profissões as quais o olhar comum não percepciona como facilmente substituíveis). O conhecimento deste contexto é especialmente relevante, exigindo tomadas de posição que sustentarão a necessidade de uma moral social actualizada.

Por fim, abordar-se-á o problema da moral computacional num contexto em que ecossistema do conhecimento estará bastante enriquecido, pois terá de incorporar agentes não-biológicos com capacidade para se tornarem intervenientes activos em dimensões que, até agora, têm sido atribuídas exclusivamente a humanos. Neste domínio serão apresentados tópicos relacionados com a Psicologia Evolucionária e com a História da Filosofia, explorando a emergência do conceito de autonomia e as virtualidades do raciocínio contra-factual e da sua aplicação no contexto da moral em IA, para darmos apenas três exemplos relevantes.

De notar que já existe em língua portuguesa vasta literatura científica em torno do tema da IA e seus afins - tome-se como exemplo *A Revolução do Algoritmo Mestre*, de Pedro Domingos, ou *A Estranha Ordem das Coisas*, de António Damásio, ou *Mentes Digitais*, de Arlindo Oliveira - todavia a aproximação às questões ligadas à moral computacional, quer na sua articulação com a moral das máquinas, quer com a moral social, não está ainda feita, nem sequer nestas obras recentes.

---

**Lista de textos recentes de Luís Moniz Pereira de base para o livro**

Estão referenciados e disponíveis via *links* na página pessoal do autor em
http://userweb.fct.unl.pt/~lmp/publications/Biblio.html

**Livro:**

L. M. Pereira, A. Saptawijaya, **Programming Machine Ethics**, Springer SAPERE series, Vol. 26, 194 pages, ISBN: 978-3-319-29353-0, DOI 10.1007/978-3-319-29354-7, Springer, **2016.**

**Capítulos de livros:**

T. A. Han, L. M. Pereira, **Evolutionary Machine Ethics**, in: O. Bendel (ed.), Handbuch Maschinenethik, Springer, **2018**.

A. Saptawijaya, L. M. Pereira, **From Logic Programming to Machine Ethics**, in: O. Bendel (ed.), Handbuch Maschinenethik, Springer, **2018**.

L. M. Pereira, A. Saptawijaya, **Counterfactuals, Logic Programming and Agent Morality**, in: R. Urbaniak, G. Payette (eds.), Applications of Formal Philosophy: The Road Less Travelled, Springer Logic, Argumentation & Reasoning series, ISBN: 978-3319585055, pp. 25-54, Springer, October **2017**.

L. M. Pereira, A. Saptawijaya, **Counterfactuals in Critical Thinking with Application to Morality**, in: Magnani, L., Casadio, C. (eds.), Model-Based Reasoning in Science and Technology: Logical, Epistemological, and Cognitive Issues, ISBN 978-3-319-38982-0, chapter DOI: 10.1007/978-3-319-38983-7_15, SAPERE series, ISSN 2192-6255, vol. 27, Springer, July **2016**.

L. M. Pereira, **Software sans Emotions but with Ethical Discernment**, in: S. Silva (ed.), Morality and Emotion: (Un)conscious Journey into Being, ISBN: 978-1-138-12130-0, pp. 83-98, , Routledge, June **2016**.

A. Saptawijaya, L. M. Pereira, **The Potential of Logic Programming as a Computational Tool to Model Morality**, in: Robert Trappl (ed.), A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations, pp. 169-210, ISBN 978-3-319-21547-1, Cognitive Technologies series, ISSN 1611-2482, Springer, December **2015.**

L. M. Pereira, A. Saptawijaya, **Bridging Two Realms of Machine Ethics**, in: J. White, R. Searl (eds.), Rethinking Machine Ethics in the Age of Ubiquitous Technology, IGI Global, ISBN13: 9781466685925, DOI: 10.4018/978-1-4666-8592-5, pp. 197-224 , July **2015**.

F. Cardoso, L. M. Pereira, **On artificial autonomy emergence -- the foothills of a challenging climb**, in: J. White, R. Searl (eds.), Rethinking Machine Ethics in the Age of Ubiquitous Technology, IGI Global, ISBN13: 9781466685925, DOI: 10.4018/978-1-4666-8592-5, pp. 51-72 , July **2015**.

L. M. Pereira, **Can we not Copy the Human Brain in the Computer?**, in: “Brain.org”, ISBN: 978-989-8380-15-9, pp. 118-126, Fundação Calouste Gulbenkian, Lisbon, **2014**.

T. A. Han, L. M. Pereira, **Intention-based Decision Making via Intention Recognition and its Applications**, in: H. Guesgen, S. Marsland (eds.), Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security, pp. 174-211, ISBN 978-1-4666-3682-8, IGI Global, **2013**.

L. M. Pereira, **Evolutionary Tolerance**, in: L. Magnani, L. Ping (eds.), Philosophy and Cognitive Science - Western & Eastern Studies. Select extended papers from the PCS2011 Intl. Conf., SAPERE series, ISSN 2192-6255, vol. 2, pp. 263-287, ISBN 978-3-642-29927-8, Springer-Verlag, **2012**.

L. M. Pereira, **Evolutionary Psychology and the Unity of Sciences - Towards an Evolutionary Epistemology**, in: O. Pombo, J. M. Torres, J. Symons, S. Rahman (eds.), Special Sciences and the Unity of Science, Series on Logic, Epistemology, and the Unity of Science, Vol.24, pp. 163-175, ISBN: 978-94-007-2029-9, Springer, **2012.**

L. M. Pereira, A. Saptawijaya, **Modelling Morality with Prospective Logic**, in: M. Anderson, S. L. Anderson (eds.), "Machine Ethics", pp. 398-421, ISBN: 978-0521112352, Cambridge University Press, **2011**.

L. M. Pereira, A. M. Pinto, **Collaborative vs. Conflicting Learning, Evolution and Argumentation**, in: H. R. Tizhoosh, M. Ventresca (eds.), Oppositional Concepts in Computational Intelligence, pp. 61-89, Springer (series Studies in Computational Intelligence 155), **2008**.

**Artigos em revistas científicas:**

L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, T. Lenaerts, **When agreement-accepting free-riders are a necessary evil for the evolution of cooperation**, ***Scientific Reports***, SREP-16-35583, DOI:10.1038/s41598-017-02625-z, online 30 May **2017**.

L. M. Pereira, **Cyberculture, Symbiosis and Syncretism**, in: ***AI & Society*** (Journal of Knowledge, Culture and Communication), DOI: 10.1007/s00146-017-0715-6, open access here, online 21 March **2017**.

A. Saptawijaya, L. M. Pereira, **Logic Programming for Modeling Morality**, in: Magnani, L., Casadio, C. (Eds.), Special Issue on "Formal Representations of Model-Based Reasoning and Abduction", of ***The Logic Journal of the IGPL***, vol. 24(4): 510-525, DOI: 10.1093/jigpal/jzw025, online 9 May, August **2016**.

T. A. Han, L. M. Pereira, T. Lenaerts, **Evolution of Commitment and Level of Participation in Public Goods Games**, in: ***Autonomous Agents and Multi-Agent Systems*** (AAMAS), DOI: 10.1007/s10458-016-9338-4, 3(31):561–583, May 2017. open access here online 14 June **2016**.

L. M. Pereira, A. Saptawijaya, **Abduction and Beyond in Logic Programming with Application to Morality**, in: Magnani, L. (Ed.), ***IfColog Journal of Logics and their Applications*****,** Special issue on *Abduction*, 3(1):37-71, May **2016**.

B. Deng, **The Robot's Dilemma**, Interviews L. M. Pereira, in: ***Nature***, pp. 24-26, vol. 53, 2 July **2015**.

L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, T. Lenaerts, **Apology and Forgiveness Evolve to Resolve Failures in Cooperative Agreements**, ***Scientific Reports***, Sci. Rep. 5:10639, DOI:10.1038/srep10639, 9 June **2015**.

T. A. Han, L. M. Pereira, F. C. Santos, T. Lenaerts, **Emergence of Cooperation via Intention Recognition, Commitment, and Apology -- A Research Summary**, ***AI Communications***, DOI:10.3233/AIC-150672, vol. 28(4):709-715, preprint online June **2015**.

T. A. Han, F. C. Santos, T. Lenaerts, L. M. Pereira, **Synergy between intention recognition and commitments in cooperation dilemmas**, ***Scientific Reports***, Sci. Rep. 5:9312, DOI:10.1038/srep09312, 20 March **2015**.

T. A. Han, L. M. Pereira, T. Lenaerts, **Avoiding or Restricting Defectors in Public Goods Games?**, ***Journal of the Royal Society Interface***, http://dx.doi.org/10.1098/rsif.2014.1203 (online: 24 December 2014), 12:103, February **2015**.

L. M. Pereira, E.-A. Dietz, S. Hölldobler, **Contextual Abductive Reasoning with Side-Effects**, ***Theory and Practice of Logic Programming***, 14(4-5):633-648, DOI: 10.1017/S1471068414000258, July **2014**.

A. Saptawijaya, L. M. Pereira, **Tabled Abduction in Logic Programs**, ***Theory and Practice of Logic Programming***, 13(4-5-Online-Supplement), July **2013**.

T. A. Han, L. M. Pereira, F. C. Santos, T. Lenaerts, **Good Agreements Make Good Friends**, ***Scientific Reports***, Sci. Rep. 3:2695, **DOI:10.1038/srep02695**, **2013**.

T. A. Han, L. M. Pereira, **Context-dependent Incremental Decision Making Srutinizing Intentions of Others via Bayesian Network Model Construction**, ***Intelligent Decision Technologies (IDT)***, 7 (4):293-317, doi:10.3233/IDT-130170, **2013.**

T. A. Han, L. M. Pereira, **State-of-the-Art of Intention Recognition and its Use in Decision Making**, ***AI Communications***, DOI: 10.3233/AIC-130559; 26 (2): 237–246, **2013**.

**A que acrescem materiais recentes, e outros que se lhes vão acrescentando na dita página:**

L. M. Pereira, F. Cardoso, **A ilusão do que conta como agente**, in: M. Curado, A. D. Pereira, A. E. Ferreira (eds.)., Vanguardas da Responsabilidade: Direito, Neurociências e Inteligência Artificial. (Col. Centro de Direito Biomédico, 26) Coimbra: Petrony, no prelo, **2019**.

L. M. Pereira, **A machine is cheaper than a human for the same task**, in: ***AI & Society*** (Journal of Knowledge, Culture and Communication), DOI: 10.1007/s00146-018-0874-0, vol. 34(1), January **2019**.

T. A. Han, L. M. Pereira, **Evolutionary Machine Ethics Synopsis**, invited paper in: ***Journal of the Japanese Society for Artificial Intelligence***, to appear in Japanese in **2019.**

# Luís Moniz Pereira’s comments on the

# EU’s “Draft Ethics Guidelines for Trustworthy AI” of 19 December 2018

https://ec.europa.eu/knowledge4policy/publication/draft-ethics-guidelines-trustworthy-ai_en

## Introduction: Rationale and Foresight of the Guidelines

1- No explicit emphasis is placed on the AI creation of wealth and its actual distribution among all humans. AI will actually ever more strongly accentuate the increasing wealth gap, unless new social compacts are put in place, there being dangerous risks of resentment and revolt otherwise, and ensuing shunning of AI, a pity because it is after all a conquest of humanity as a whole. The whole question of societal wealth and values is being given short shrift or swiped under the rug.

2- Machines, whether robots or software and their combination, will themselves have to act morally to be convivial with us (and amongst themselves). But we know too little about our own ethics and how to impart it to machines. More ethics research is required, starting now.

3- Similarly, more jurisprudential conceptual scaffolding is needed that will support laws, regulations and standards, including the use of LAWS (Legal Autonomous Weapon Systems) and autonomous machines in general.

4- The Guidelines should foresee regulations and monitoring concerning the activity of contract consortia, such that individual responsibility is clearly defined from the start -- the so-called "Problem of Many Hands."

5- Joint EU initiatives such as CLAIRE, and international collaboration centres (viz. CERN), should be spelled out as natural venues for increased and widespread value of AI, at the same time striving to avoid the most pernicious dangerous aspects of an AI race, by joint validation, certification, monitoring, and agreed joint AI security.

6- International rules of commitment should be fostered, subscribed and monitored, like with climate change agreements.

## Chapter I: Respecting Fundamental Rights, Principles and Values - Ethical Purpose

1- The issue of societal values concerning wealth distribution is skimmed over in this chapter. AI will increasingly and acutely widen the pre-existing and wealth gap already on the increase. Not enough concern is shown in the Guidelines regarding the unstoppable encroaching of machines into the heretofore human monopoly of cognition and hand-eye coordination, and overall negative impact on unemployment. The immense technical progress brought about by AI is not being accompanied by a concomitant social progress that will benefit everyone's actual wealth and less striving for a living, not just for the owners of patrimony and technology.

2- The old capital/labour split needs urgent revision. After all, my body is my own limited capital, so even after I leave a company for another, the body capital I spent in the first should continue to benefit me thereafter if that company is successful.

## Chapter II: Realising Trustworthy AI

1- Computer languages need to be developed that enable the specification, validation and monitoring of ethical constraints in programs.

2- Programmed AI machines must be subject to safety and compliance tests before being marketed. A case in point are driverless cars, which must comply with common standards imposed by authorities, who thereby become jointly responsible for untoward incidents as a result of improper certification.

3- A recent law that went into effect in California already in 2019, prohibits software that impersonates a human. That should be easy to rapidly obtain consensus on.

4- Large windfall profits should commit to a margin to help promote trustworthy AI by independent organisations.

## Chapter III: Assessing Trustworthy AI

1- International chartered bodies are needed to enact and assess the trustworthiness of AI and be enabled to denounce violations.

2- Independent and credited auditors must be set up, over and above internal auditing by companies, governments, and protected individual denouncing of risks.

## General Comments

1- Stakeholders must include the Humanities, since the impact of AI is quite wide and needs contributions from a diversity of fields of knowledge, that must be promoted to best contribute. Specifically, I point out Philosophy, Psychology, Ethics, Jurisprudence, Linguistics, Anthropology, Sociology, Economics, Political Sciences, Evolutionary Science.

2- AI research, largely construed, should be further concentrated, centred and promoted in the universities (and research institutes), and there it can easier and more naturally be interdisciplinary in character.

3- A tax on sales is needed, over and above that on profits (always hard to audit because of globalisation and fiscal paradises).

4- A tax on robots and software fully replacing humans must be contemplated, for replacing means replacing, including social security contributions by the worker and the employer. That will help prevent social disruptions.