17

Luís Moniz Pereira

“On Morals for Machines and The Machinery of Morals”

CONFERÊNCIAS

HORIZONTES DO FUTURO

loulé concelho

Loulé Cidade Educadora

LOULÉ 2019

CONFERÊNCIAS

HORIZONTES DO FUTURO

LUIS MONIZ PEREIRA

"Moral para Máquinas & a Maquinaria da Moral"

Quinta-Feira, 21 fevereiro 2019 | 21h30
Salão Nobre da Câmara Municipal de Loulé

loulé concelho

WWW.CM-LOULE.PT

Ficha Técnica:
Autor: **Luís Moniz Pereira**
Título: **"On Morals for Machines & The Machinery of Morals"**
Coleção: *Caderno Conferências Horizontes do Futuro*
N.º 17, Abril 2019
Editor: Câmara Municipal de Loulé
Impressão: Gráfica Comercial de Loulé
Depósito Legal: 455193/19
ISBN: 978-989-8978-00-4

ON MORALS FOR MACHINES

&

THE MACHINERY OF MORALS

Luís Moniz Pereira

NOVA LINCS – Universidade Nova de Lisboa

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

CÂMARA MUNICIPAL DE LOULÉ, 21st February 2019

ABSTRACT

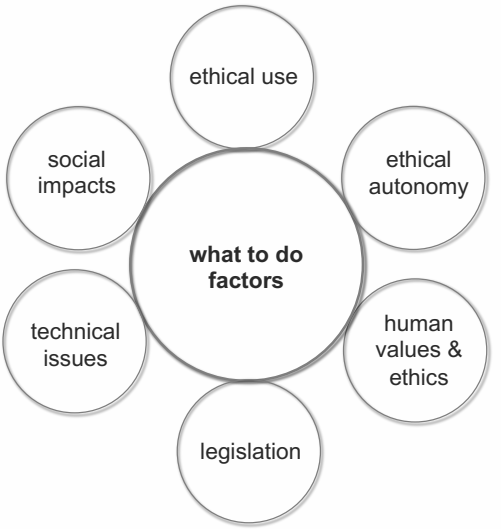
• We are at a crossroads between **Artificial Intelligence, Machine Ethics**, and their **Social Impacts.**

• 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

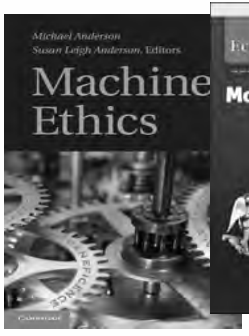


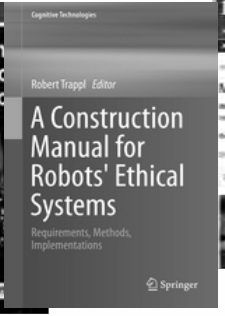
Ethical machines – the why and the how

- **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 greatattention, 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 - 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.

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 dualprocess 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.

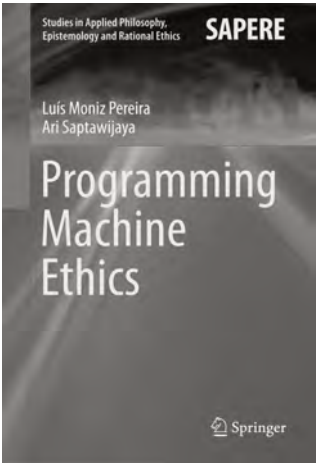
Programming Machine Ethics

• 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.



This more technical book of mine addresses such issues.

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.

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 oberable 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.

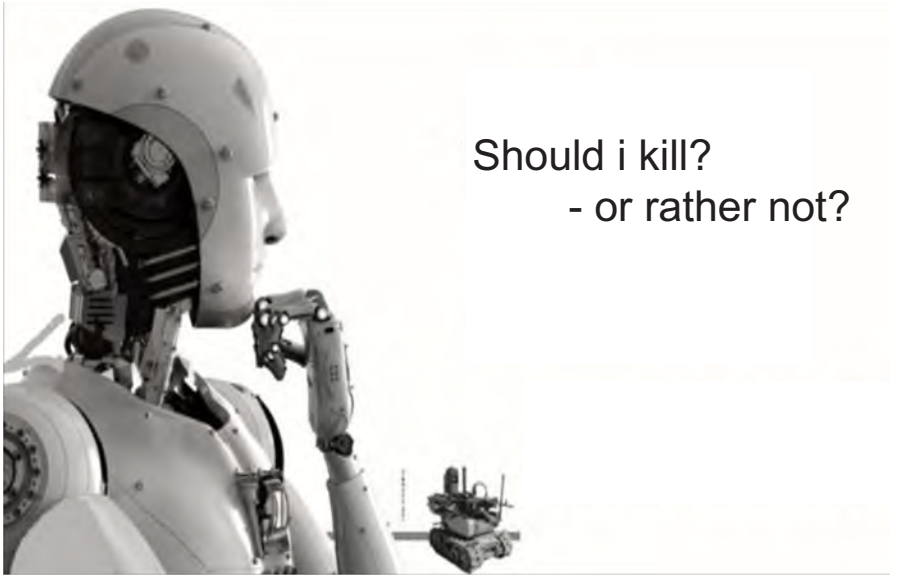
• We know not yet nearly enough about human morality.

• 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!



Thanks for your attention

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

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.1

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.

Many thanks to my co-authors:

- Ari Saptawijaya (Indonesia)
- The Anh Han (UK)
- Tom Lenaerts (Belgium)
- Francisco C. Santos (Portugal)
- Luis Martinez-Vaquero (Italy)

Da moral da máquina à maquinaria da moral

Luís Moniz Pereira

Palestra na Câmara de Loulé, 21 de Fevereiro 2019

O título diz quase tudo, i.e., vamos endereçar o problema de criar uma moral da máquina, ou seja, da necessidade de as máquinas terem uma moral. Em resumo, as máquinas estão cada vez mais sofisticadas e autónomas, têm que coabitar connosco, conviver connosco e, portanto, têm que se inserir na nossa própria moral, que é aquilo que nos agrega como sociedade gregária que somos.

A maquinaria da moral, que também está no título, visa chamar a atenção para que no fundo a moral é uma série de regras que são mecanismos, que são maquinaria no sentido de que há mecanismos morais que as sociedades adoptam. Pelo facto de serem mecanismos, e se nós os compreendermos bem, estarão próximos de serem postos na máquina, como mecanismos que são. Portanto o problema será o de como dar moral às máquinas começando por perceber a nossa moral como uma colecção de mecanismos, de modo a que tal nos ajude então a vir a programar as máquinas com a nossa moral.

O assunto é vasto, tem múltiplas fronteiras com muitas ciências, com a Filosofia, com a Psicologia, etc., como veremos. É tão vasto que farei um esforço agora de dar apenas algumas pinceladas largas de modo a cobrir as suas principais dimensões, fugindo à tentação de entrar demasiado nos detalhes desde já.
Em síntese, estamos numa encruzilhada, a da IA, da ética das máquinas, e do seu impacte social. É uma situação nova porque pela primeira vez vamos ter seres, que não somos nós, mas que vão conviver connosco. Seres esses que terão um impacto significativo nas nossas vidas, nas próximas dezenas de anos para não dizer desde hoje em dia, assim como na nossa sociedade.

O tópico da moral tem dois grandes domínios, que investiguei e em que trabalho. Um, a que chamo do "cognitivo", isto é, o de como é que pensamos em termos morais.
Para ter um comportamento moral é preciso equacionar possibilidades: devo eu comportar-me assim ou comportar-me assado? É preciso equacionar os vários cenários, as várias hipóteses. É preciso comparar essas hipóteses para ver quais são as mais desejáveis, quais são as suas consequências, quais são os seus efeitos laterais. Tudo isso envolve ter capacidades cognitivas para o fazer, como seja prospectar o futuro, ser capaz de olhar para o futuro.
O nosso cérebro tem essa capacidade, que lhe é essencial para a vivência em sociedade, e justamente as referidas capacidades cognitivas têm que ser úteis para a vivência em sociedade, quer dizer, a moral não é uma coisa individual, algo que alguém isolado numa ilha precise; a moral é necessária numa população para que esta coopere, seja gregária e consiga comportar-se de uma maneira vantajosa para todos.
Esse é o outro domínio. E é esse o problema da moral, o de garantir a vantagem comum, em vez de cada um só fazer por si.

Portanto, a existência de comportamentos morais numa população exige certas capacidades cognitivas, e as que temos também determinam o que é possível em convivência, no tal domínio societal ou populacional. Por exemplo, uma capacidade cognitiva outra, além da que mencionei de olhar para o futuro, é a de olhar para o passado, ou seja, a de ser capaz de pensar, sabendo o que sei hoje, o que teria feito ao invés na altura. E posso usar isso para dar recomendações a pessoas que hoje se encontrem na situação em que eu estava na altura. Isso permite uma certa aprendizagem social, mas exige essa capacidade cognitiva de imaginar como é que o passado podia ter sido diferente. É mais um exemplo de capacidade cognitiva a propósito da moral.

Na minha investigação tive que estudar certas capacidades cognitivas para depois ver ser elas eram promotoras de cooperação, numa população de seres informáticos. Ou seja, de programas que convivem entre si, em que um programa é um conjunto de estratégias definidas por regras. Isto é, numa dada situação um

programa tem uma certa acção ditada pela sua estratégia, e os outros programas têm igualmente acções ditadas pelas suas. É como se fossem agentes convivendo em conjunto, cada um com estratégias possivelmente diferentes. Estuda-se o se, e o como essa população vai evoluir num bom sentido, e se esse sentido é estável, se se mantém no tempo.

Um instrumento muito importante para o fazer é a chamada Teoria dos Jogos Evolucionários (Evolutionary Game Theory - EGT), que consiste em ver como é que num dado jogo uma população evolui por aprendizagem social. Afinal de contas a sociedade rege-se por um conjunto de regras de funcionamento conjunto, digamos regras dum jogo em que se permite fazer certas coisas e não outras.

O jogo indica quais os ganhos ou as perdas de cada jogador em cada jogada, consoante o modo como joga. A aprendizagem social consiste em um jogador passar a imitar a estratégia de um outro jogador cujos resultados indicam ter maior sucesso.

Dadas certas regras, como é que evolui esse jogo social? Aqui poderíamos entrar pelo campo da ideologia, mas não vamos tão longe. Estamos ainda a estudar como é possível a moral. Porque supomos que a moral é evolucionária, que evoluiu com a nossa espécie.

À medida que fomos evoluindo, ao longo de milhões de anos, fomos aperfeiçoando as regras de convívio e aperfeiçoando as nossas próprias capacidades intelectuais de saber usar regras de convívio. Nem sempre bem, e isso é um problema constante: ou seja, as regras sociais serem tais que todos beneficiemos delas, embora haja sempre a tentação de uns quererem beneficiar mais do que os outros, de obterem as vantagens sem pagar os custos.

É esse o problema essencial da cooperação: como é que esta é possível mantendo sob controlo os que querem abusar dela. Para a nossa espécie chegar onde chegou hoje, a própria evolução teve que nos ir selecionando em termos de uma moral de convivência gregariamente proveitosa. Voltaremos recorrentemente a essa problemática.

Queria enfatizar, no entanto, que estamos perante uma Terra Incognita, que há todo um continente por explorar, de que estamos apenas a aflorar os contornos. Não sabemos ainda muito sobre a nossa própria moral, e de forma suficientemente exacta para a

podermos programar em máquinas.
Na verdade, há várias teorias éticas, competidoras entre si, mas também complementando-se. A filosofia e a jurisprudência estudam a Ética, que é a problemática de definir um sistema de valores articulado em princípios. Cada ética, em particular, é o substracto de suporte às normas e legislação que justificam, em cada contexto, as regras específicas vão aplicar e usar no terreno essa ética, tendo como resultado regras morais concretas, dependendo das culturas e das circunstâncias.
Parte-se de princípios éticos abstractos para regras morais concretas para cada contexto. Na prática, uma moralidade, um conjunto de regras morais, resulta de uma combinação histórica, contextual e filosófica de teorias éticas que foram evoluindo no tempo.
Quando não for importante a distinção, empregarei indiferentemente ambos os termos, "ética" e "moral", como é uso comum.

Numa minha página (em https://userweb.fct.unl.pt//~lmp/publications/Biblio.html) podem consultar-se dezenas de artigos meus de investigação de carácter técnico sobre ética/moral em ambos os domínios, o do cognitivo e o do populacional. É uma investigação baseada em teorização, programação, experimentação, e verificação de consonância interdisciplinar com o que se conhece da realidade, evolutiva e presente.

O carrossel anterior (ver pág. 4) resume de certa maneira a complexidade da problemática da maquinaria moral. No seu centro nós queremos identificar aqueles factores que dizem o que fazer, o como agir. "O que fazer" está rodeado de outros tantos carrosséis, recursivamente, se assim o entendermos.
Um tem a ver com o uso ético das máquinas. Já ouvimos falar de fake-news e de algoritmos que influenciam eleições. Um mau uso das máquinas que deve ser susceptível de regras morais. Por exemplo, é um uso mau, imoral, um programa fazer-se passar por um humano. No dia 1 de Janeiro de 2019, se a data não me falha, entrou em vigor na Califórnia uma lei que diz ser proibido, um computador, um programa, fingir que é ser humano. Claro que há outros exemplos de usos imorais: os drones com capacidade

autónoma que matam pessoas.

Mas o uso imoral cada vez depende mais da própria máquina justamente porque ela tem cada vez mais autonomia, e as questões do uso moral por consequência ampliam-se.

Poderíamos por isso pensar que as máquinas nos deveriam também defender do seu uso não ético por parte de um humano. Suponhamos que alguém comandava um programa dando-lhe uma instrução para agir de um modo que causaria prejuízo a seres humanos. O próprio programa poderia recusar-se a fazer isso.

Nós temos hoje em dia programas que controlam aviões, barcos, comboios TGV, e os informáticos podem provar que um dado programa está correcto. Por exemplo, que o programa do barco nunca vai tentar fazer o barco saltar uma ponte (...), ou que o programa do comboio não irá fazê-lo andar depressa demais numa curva.

Põe-se a mesma questão de prova de correcção em relação a máquinas programadas autónomas, e mesmo em relação às que, não sendo autónomas mas controladas por um humano. De modo a terem a capacidade de dizer “não, eu não faço isso.” E nós sermos capazes de provar, com técnicas da Informática, que aquele programa nunca irá agredir um ser humano, ou nunca se irá querer passar por um ser humano, fingindo que o é.

É essa também uma razão pela qual nós precisamos de introduzir moral nas máquinas, para que elas não façam tudo o que lhes dizem, ou seja, não queremos que a máquina esteja na situação de dizer “fiz isso porque me mandaram.” Essa foi a posição dos criminosos nazis em Nuremberga, dizendo “eu fiz isso, mas foi só porque me mandaram,” como se não tivessem sentido crítico e não soubessem desobedecer a ordens. Nós temos que saber construir máquinas que saibam desobedecer a certas ordens.

Outra plataforma do carrossel acima é a dos Valores Humanos. Nós no fundo pretendemos dar às máquinas os nossos valores, porque elas vão conviver connosco. Claro que se mandarmos para Marte uma troupe de máquinas elas poderão ter a sua própria moral, apropriada ao ambiente e à tarefa, não havendo lá humanos. As máquinas que conviverem connosco, no entanto, têm de estar moralmente conciliadas com a população onde se

encontram.

Noutro círculo do carrossel indico a Legislação porque, no fim do dia, tudo vai ter que se traduzir em leis, normas e standards, sobre o que é permitido ou proibido.

Tal como os carros têm standards de poluição também as máquinas terão que obedecer a certos standards, e é importante saber que um carro sem condutor obedece a standards aprovados por uma entidade com a capacidade para o fazer, como seja um órgão governamental, e se possível aprovados por uma entidade internacional.

Muitas vezes pergunta-se quem é responsável por um carro sem condutor atropelar um peão quando podia não o ter feito? É o dono, é o fabricante? Mas nunca se fala no Legislador. Contudo, alguém teve que dizer "este carro sem condutor pode circular." Tal como nós temos que tirar a carta de condução, também os carros sem condutor vão ter que tirar uma carta especialmente feita para eles. Terá que ser um governo a legislar quais os testes que um carro sem condutor deve passar. Ao verificar-se que tais testes não foram suficientemente seguros, a entidade que autorizou a circulação daqueles carros também é responsável!

Outro círculo é o das Questões Técnicas. Tudo isto envolve sempre a parte de realização técnica das máquinas, para o que que quer que seja. Nem tudo é tecnicamente possível. Por exemplo, ainda não sabemos fazer a tal prova de que uma máquina não vai fazer coisas eticamente erradas. Já para não falar em que um hacker possa entrar na máquina e obrigá-la a fazer coisas erradas. É um problema de segurança que tem de ser resolvido de um modo técnico.

Por fim, e não menos importante, são os Impactes Sociais das máquinas com autonomia. Quando me refiro a máquinas estou a falar quer em robôs quer em software. Este último é bem mais perigoso pois espalha-se e reproduz-se facilmente em qualquer lugar do mundo, ao passo que o robô é muito mais difícil de reproduzir: tem um custo maior e tem limitações físicas.

No que toca ao impacte social, embora às tantas iremos ter robôs a cozinhar hambúrgueres e a servir-nos à mesa, para o que não é

preciso ser muito inteligente, mas é antes do mais preciso ter uma coordenação fina olho-cérebro-mão. Algo que as máquinas ainda não têm tanto quanto os humanos, mas também nessa frente os robôs estão a avançar muito depressa.
No que toca ao software, a questão é mais preocupante porque no fundo os programas estão a entrar em níveis cognitivos que até agora era nosso monopólio. Daí que as pessoas se sintam muito preocupadas. Quer dizer, havia até agora coisas que só um humano sabia fazer, mas pouco a pouco a pouco as máquinas começaram a jogar xadrez, a fazer diagnóstico médico, etc., e cada vez irão fazer actividades mentais mais sofisticadas e irão aprendendo a fazê-lo sucessivamente.
Essa porta que se abriu, cria pela primeira vez uma competição com os humanos que pode fazer com que, dependendo da organização social, da ideologia, e da política, os humanos vão sendo substituídos por máquinas, porque para fazer a mesma coisa a máquina sai mais barata do que o humano. Então sendo o humano dispensável, os salários irão diminuir e os donos das máquinas irão ser cada mais ricos.
O presente hiato, que vem aumentando e que indica que os ricos estão cada vez mais ricos e os pobres cada vez mais pobres, é um hiato que a IA está já a, e virá ampliar ainda mais. A tal ponto que exigirá um novo contrato social, sob pena de um cataclismo. A maneira com funcionamos em termos de capital e trabalho, e o como se equacionam as duas coisas, tem que ser completamente reformulada. Com o risco de, se tal não acontecer, os hiatos de riqueza vão fazer com que, mais tarde ou mais cedo, ocorra uma grande revolta, sublevação, e desagregação social. Não será como o actual ressentimento dos coletes amarelos (gilets jaunes), mas muito mais ampla e profunda do que isso. Uma sublevação talvez até mesmo maior do que, ou pelo menos comparável, à da Revolução Francesa. Vai aparecer uma grande revolta quando o sistema de castas, entretanto instalado, rebentar.
Para evitar esse grande incêndio social, urge desde já começar a limpar as bermas da sua propagação, e iniciar a estruturação dum novo contrato social.

Actualmente já temos robôs em hospitais, temos drones que voam, temos lanchas aquáticas, temos carros sem condutor, e temos até jogos morais interactivos, susceptíveis de ensinar moral. Num jogo que fiz, um robô vai salvar uma princesa, para o que combina várias aproximações éticas.[1] Inclusivamente tal exemplifica-nos que a moral na prática não é só uma, mas, mais vulgarmente, uma combinação delas. Nós próprios não seguimos a moral do cavaleiro andante, ou a moral utilitária, ou a moral kantiana, ou a moral do Gandhi. A nossa moral é uma mistura delas e vai evoluindo. Nesse programa-jogo mostro como é que a moral do robô vai evoluindo. Temos, pois, que assumir que a programação da moral nas máquinas tem de vir a permitir a sua própria evolução.

Não há uma moral fixa, congelada. A moral é uma coisa evolutiva, e ao longo da história da espécie, remota e próxima, tem vindo a evoluir. Interessa relevar que as lanchas voadoras aquáticas, tal como os drones o podem fazer, coordenam-se em enxames delas para atacar navios inimigos. Não estamos, assim, a falar apenas de uma moral do indivíduo isolado, mas de uma moral em que as máquinas, ao agirem em conjunto, resulta daí a eclosão de um comportamento distribuído, eventualmente imprevisto.
Por exemplo, podemos imaginar um enxame de drones numa fronteira, a controlar os movimentos de imigrantes, e a tentarem afastá-los dos poços de água e dos bons caminhos, ou assustá-los de alguma maneira. E a dada altura alguém há de disparar um tiro para um drone um drone irá disparar um tiro para baixo. Actualmente já há drones que agem em pelotões, o que torna muito mais difícil controlar o seu comportamento. Este já não é previsível a partir do drone individual, mas é o resultado emergente de uma população de drones. Daí o ser tão importante estudar a moral em termos de populações e dos seus parâmetros de configuração.

[1] Pode ser visto aqui: https://drive.google.com/file/d/0B9QirqaWp7gPUXBpbmtDYzJpbTQ/view?usp=sharing, sendo também explicado em detalhe, em inglês, aqui, bem como nas referências lá constantes: https://userweb.fct.unl.pt//~lmp/publications/online-papers/lp_app_mach_ethics.pdf. O robô vai mostrando num balão o que está a pensar, e mostra-se como o utilizador lhe vai dando novas regras morais a juntar às anteriores, por vezes suplantando-as quando há contradição entre elas.

Está claro que as máquinas estão cada vez mais autónomas e nós temos que garantir que elas conviverão connosco nos termos das nossas regras. Há, portanto, um novo paradigma moral que diz que a moral também é computacional. Quer dizer, temos que saber capazes de programar a mora. Isto tem um lado positivo, porque ao programarmos a moral nós percebemos melhor a nossa própria moral.

Dou-vos um exemplo: tenho um trabalho científico sobre culpa, no qual introduzo a culpa em populações de agentes informáticos. Passam a ter capacidade de culpa e as sentirem-se culpados quando fazem algo que prejudica outro, resultando numa espécie de autopunição, e mudança de comportamento, que evitará culpabilizações futuras. Não é uma culpa no sentido existencial, freudiano, mas no aspecto pragmático de não ficarem contentes com o que fizeram ao prejudicar alguém.

Introduza-se uma dose de culpa, nem de mais nem de menos, em apenas alguns agentes duma população deles que interage dentro do computador num jogo evolutivo. Sem a existência de culpa, a maioria tende a jogar egoisticamente, cada um querendo ganhar mais do que os outros, não se conseguindo por isso chegar ao nível em que todos podem ainda ganhar mais. Mas este resultado desejável já é tornado possível se houver uma dose de culpabilização inicial, que modifica comportamentos e se espalha como boa estratégia a toda a população.

Nós conseguimos mostrar matematicamente que uma certa dose de emoção de culpa é vantajosa e promove a cooperação. Não pode ser nem excessiva nem diminuta. E também que não se deve sentir culpa face a quem não se sente por sua vez culpado, pois tal é deixar-se ser abusado.[2]

Este é aliás o grande problema abstracto central da moral e do gregarismo, que naturalmente incide também no caso nas máquinas: Como conseguir evitar o egoísmo puro dos agentes que

[2] Para os detalhes técnicos consulte-se: L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, T. A. Han, **Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems**, in: Procs. 16th Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2017), Das, S. et al. (Eds.), pp. 1422-1430, 8–12 May 2017, São Paulo, Brazil.

oportunisticamente se querem aproveitar do gregarismo dos outros sem quererem, por sua vez, contribuir para ele?
Por outras palavras, como poderemos demonstrar, através de modelos matemáticos computacionais, em que circunstâncias o gregarismo é evolucionariamente possível, estável e vantajoso?
Podemos usar o próprio computador para percebermos melhor como é que a maquinaria da culpa funciona, entre que valores de parâmetros, e variarmos esses parâmetros para perceber como melhor os usar evolutivamente.

E a dada altura, ao criarmos agentes artificiais que têm culpa, uma certa dose de culpa, damos ao mesmo tempo argumentos ao facto de a culpa ser uma função útil, resultado da nossa evolução. Ou seja, porque a culpa é útil, e como é útil fomos selecionados, ao longo da evolução, a ser capazes de a ter. E a ser capazes também de induzir culpa nos outros.
Contribuindo até para explicar o termos uma religião católica muito baseada na ideia de culpa: A pessoa já nasce culpada, já nasce a dever qualquer coisa. E conseguirmos começar a perceber qual é o papel computacional de certas facetas morais embebidas no nosso próprio sistema nervoso. No fundo, são facetas “compiladas” na espécie, para usar uma expressão informática.

Bom, como já se percebeu isto é um tema que vive paredes meias com a Filosofia, com a Jurisprudência, com a Psicologia, com a Antropologia, com a Economia, etc., em que são importantes a interdisciplinaridade e a inspiração que nos dão esses vários domínios.

Embora aproveite para chamar a atenção que um dos problemas que temos é o de que a Jurisprudência não está a avançar suficientemente tendo em vista o legislar sobre as máquinas morais. Porque, por exemplo, há vários tipos de autonomia de máquinas, mas as nossas leis são feitas para seres humanos, que pressupomos terem uma certa autonomia tipo, a não ser que estejam doentes ou doidos. Quando fizermos legislação com respeito às máquinas, temos que definir e começar por usar conceitos sem os quais será impossível fazer leis, pois estas têm sempre que apelar aos conceitos da jurisprudência.

É importante reconhecer que o Legislador está muito atrasado em acompanhar o passo da técnica, e isso é preocupante porque há uma grande confusão entre a noção de que o progresso técnico é igual ao progresso social. Na verdade, o progresso técnico, a que assistimos aceleradamente por todo o lado, não significa que esteja a ser acompanhado por um desejável e concomitante progresso social. A técnica deve estar a ser usada ao serviço dos valores humanos, e esses valores devem ser usufruídos igualmente por todos, com a criação de riqueza a ser distribuída com justiça.

Gosto de dar a analogia de que o grande progresso e apogeu na civilização Grega, séculos V e IV a.C., foi possível porque tinham escravos, sem direitos de cidadania nem possibilidade de ascensão, e que eram essencialmente constituídos pelos exércitos que conquistavam, e cidadãos estrangeiros.

Ora nós temos a possibilidade de usufruirmos cada vez mais de máquinas escravas, que já o são, a libertarem-nos do esforço que pode ser feito por elas.

Mas gostaríamos que toda a gente fosse liberta e ganhasse igualmente com isso, através de uma distribuição justa da riqueza produzida por tais escravos, que são desta feita naturais.

E não, pelo contrário, que essas máquinas substituam pessoas resultando daí um lucro cada vez maior para os donos exclusivos das máquinas. E não, também, que as pessoas fiquem desempregadas ou com salários mais baixos, ao competirem com máquinas com as quais será cada vez mais impossível competir.

Daí ser indispensável um novo contrato social, em que a relação trabalho/capital seja reformulada e actualizada, em consequência do impacte social das novas tecnologias, nomeadamente de sofisticadas máquinas com cognição e autonomia.

Costumo dizer: Se uma máquina me vai substituir deve fazê-lo completamente. Neste sentido: Eu, ao posicionar-me numa actividade de trabalho, contribuo para a segurança social que sustenta os reformados actuais; contribuo para o Serviço Nacional de Saúde; contribuo com o IRS para tornar possível a governação e desenvolvimento do país, etc.

Consequentemente, se uma máquina me substituir completamente,

eliminando uma pessoa de um trabalho cuja actividade se mantém, também deve pagar os impostos que eu estava pagando para sustentar o contrato social vigente. Substituir é substituir! É substituir nesses aspectos todos.

Se estivéssemos a falar de engenharia civil, diríamos claro que os engenheiros civis se preocupam com a segurança e a qualidade. Há normas, regras, para um edifício resistir aos tremores de terra, para as paredes isolarem o ruído, etc. Não me competirá preocupar-me com o impacte social disso. E muito bem, só que não é possível reduzir os problemas da ética das máquinas a um código deontológico que os engenheiros informáticos devam seguir. Justamente pelo impacto que isso tem nos valores humanos e organização social, e no nosso devir civilizacional. Por isso a questão dos valores é ineludível, e não se pode reduzir a standards técnicos.

Efectivamente, há vários e repetidos relatórios de estudos, compagináveis, de entidades insuspeitas, nomeadamente da McKinsey & Company, do Pew Research Center, da OCDE, da PricewaterhouseCoopers, etc., que apontam para um acréscimo de entre 15-20% de desemprego adicional em 2030, só em virtude da IA.

Na China será pior, será mesmo de 20%, porque enquanto no mundo ocidental as pessoas ainda podem trepar um pouco, tornar-se especializadas cognitivamente dado o seu mais alto ponto de partida educativo, na China o nível de educação parte de mais baixo e, portanto, a capacidade de as pessoas subirem nas suas capacidades cognitivas e manterem-se à frente das máquinas é mais baixa e lenta. Imaginem, pois, a massa empregável de 1,4 mil milhões de chineses e no respectivo impacte.

O tópico do desemprego causado pela IA nas próprias superpotências de IA, e que noutras paragens será mais gravoso, é bem analisado no recente livro de Kai-Fu Lee.[3]

Aí está o futuro: Once upon a time apareceu uma sociedade de castas: a dos donos dos robôs, a dos managers das máquinas,

[3] Kai-Fu Lee, "AI super-powers – China, Silicon Valley, and the New World Order", New York: Houghton Mifflin Harcourt, 2018.

a daqueles que treinam as máquinas, e a dos restantes. Neste momento há médicos a treinarem máquinas a lerem raios-X, e mais isto e mais aquilo. Está a acontecer por todo, o mundo seres humanos a ensinar as máquinas que os vão substituir. As pessoas estão a ensinar quem as vais substituir. Esta sociedade de casta a dada altura vai explodir, as pessoas já não vão aguentar mais tanta hipocrisia, tanta falta de distribuição de riqueza, tanta mentira automatizada com as fake news, e por aí fora.

Os perigos da IA não é se vai aparecer um Exterminador. Os perigos são que neste momento estão máquinas simplistas a tomar decisões que nos afectam, mas que por lhes chamarmos "máquinas inteligentes" as pessoas acham que estão a fazer um bom trabalho. Este excesso de venda da IA que ocorre actualmente é muito pernicioso nesse aspecto.
Até porque a IA que está a ser vendida não chega a um décimo do que é a IA. A IA a sério ainda está por vir, que é muito mais sofisticada do que a generalidade dos programas actuais, ditos de deep learning. Como digo, estes são coisas muito simples e não se pode estar a dar tanto poder a máquinas tão simples. Mas como substituem humanos, como vão substituir radiologistas, condutores de automóveis, camiões, pessoas nos call-centers, pessoas na segurança dos centros comerciais, são vendidos como uma panaceia.

Faço parte de um projecto patrocinado pelo Future of Life Institute (FLI), uma organização sem fins lucrativos endossada por pessoas como Stephen Hawking e Elon Musk, entre outros. O projecto é sobre o problema de que a pressa em chegar ao mercado, por parte das firmas que desenvolvem produtos de IA, ir fazer com que elas descurem as condições de segurança desses produtos. A pressa é tal que a segurança é posta de lado porque custa dinheiro e tempo, e atrasa a chegada ao mercado antes dos competidores.
O projecto é sobre como é que podemos estabelecer as regras do jogo de forma a que não haja ninguém a fazer olhos mortos em

relação à segurança, como se ela não fosse essencial.[4]

Para isso precisamos de entidades que regulem e monitorizem. Acho que deveria haver uma "Comissão Nacional de Ética para a IA," que inclua a Robótica. Tal como há uma "Comissão Nacional de Bioética," ela terá que ser independente e estar acima de outras, e responder directamente ao Presidente da Assembleia da República, sem dependência do Governo.

Não podemos aceitar, como se ouve dizer na Europa e nos EUA, que as firmas que fazem carros sem condutor é que são completamente responsáveis, e que e houver algum problema logo se vê. Os Governos não são, portanto, responsáveis pelos testes a que tais carros devam ser submetidos, simplesmente delegam nas próprias firmas. Mas atente-se aos recentes acidentes ocorridos com o Boeing 737-M, em que a Federal Aviation Authority (FAA) americana delegou na própria Boeing a verificação de segurança!

Na União Europeia a responsabilidade pela segurança poderá ser mais disfarçada. Criou-se uma comissão de alto nível para a IA e Ética para dar recomendações, e o que se propõem fazer, segundo julgo, é afirmar: Temos aqui umas recomendações que as firmas devem seguir. Temos além disso aqui umas firmas privadas de auditoria que vai auditar essas firmas.
Iremos, porventura, cair no mesmo esquema de auditores interessados nas próprias entidades que estão a auditar, por que estas em paralelo lhes encomendam estudos. Veja-se o caso dos bancos e da crise financeira de 2008.

Daí a necessidade de haver alguma entidade reguladora independente, não privada.

Enfim, vivemos nesta sociedade cada vez mais algorítmica, com tudo cada vez mais sistematizado, em que um perigo crescente é o de darmos excessivo poder a máquinas simplistas, pelo risco que existe de elas cada vez mais sistematicamente nos meterem em gavetas. Porque estas máquinas simplistas, de Deep Learning sobre Big Data, o que fazem no fundo é reconhecer padrões

[4] Para uma súmula do projecto veja-se: T. A. Han, L. M. Pereira, T. Lenaerts, **Modelling and Influencing the AI Bidding War: A Research Agenda**, in proceedings of: AAAAI/ACM Conference on AI, Ethics, and Society, (AIES 2019), January 27-28, 2019, Honolulu, Hawaii, USA.

específicos num universo de padrões possíveis. Para certas coisas é óptimo, é uma técnica excelente. Não se pode é julgar que vão resolver problemas para os quais essa técnica não é adequada.

Gosto de dar este exemplo a propósito. Nos EUA há, que eu saiba, 3 programas usados pelos juízes que têm de decidir se a um dado prisioneiro é dada liberdade condicional ou não. Como funciona? Os juízes estão muito ocupados pois há milhões de prisoneiros (julgo que 0.6% da população está em prisões). Então vão ver a um historial Big Data as pessoas que tiveram liberdade condicional, e se correu bem ou não. À sua frente têm um candidato cujo perfil tem um dado padrão, consoante a idade, etnia, religião, zona geográfica, etc.

O sistema informático, em fracções de segundo, diz ao juiz em que gaveta padrão entra o perfil do candidato, e tal determina a decisão do juiz, de forma rápida e barata.

Cada caso não é um caso. Os prisioneiros são metidos em gavetas estatísticas, que pressupõem que o passado é igual ao futuro. Que a população com um certo perfil não evoluiu, nem os costumes sociais. As pessoas são julgadas pelo padrão histórico, e o vai além disso confirmar, pela inclusão de mais uma instância.

Este é um exemplo de mau uso, efectivo e real, destes algoritmos simplistas. Não quer dizer que tais algoritmos não tenham o seu nicho próprio de grande utilidade, onde são técnicas muito boas nesse nicho. Em sua defesa neste caso de aplicação, argumenta-se que os juízes, atarefados como estão, e sobretudo depois do almoço, decidem pior!

Mas esse é um problema diferente, e que também ocorre com os médicos. À medida que são pressionados para ver mais doentes por hora são obrigados a recorrer a semelhantes programas inteligentes detectores de padrões, sem espaço para exercerem sentido crítico com seu conhecimento específico, que se deseja actualizado, e caso lhes chamassem a atenção para as limitações desses programas.

Em jeito de conclusão, queria reforçar que:

- Precisamos saber mais sobre as nossas próprias facetas morais para conseguirmos passá-las às máquinas. Contudo não sabemos ainda o suficiente sobre a moralidade humana. Nesse sentido, é importante reforçar o estudo dela pelas humanidades e ciências sociais.

- A moralidade não é apenas acerca de evitar o mal, mas também acerca de como produzir o bem. Maior bem para maior número de pessoas. A problemática do desemprego é inerente a este ponto de vista.

- As universidades são o sítio apropriado para todas estas questões, pelo seu espírito de independência, a sua prática de raciocínio e discussão. E contêm nas suas faculdades a necessária interdisciplinaridade.

- Tão cedo não vamos ter máquinas com uma capacidade moral geral. Teremos máquinas que sabem respeitar normas num hospital, numa prisão, e até as normas de guerra. Estas são até as mais bem especificadas e também subscritas por todo o mundo. Como estão bem especificadas são menos ambíguas e mais próximo de poderem ser programadas.

- Iremos começar por automatizar normas e suas excepções, pouco a pouco alargando a generalidade e a capacidade de uma máquina aprender novas normas, e de ampliar os seus domínios de competência.

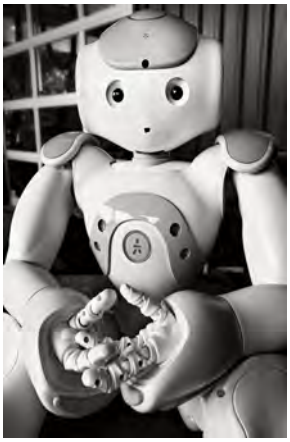
- Como são assuntos muito difíceis quanto mais cedo começarmos melhor!

FIM

O meu próximo livro intitula-se
“Moral da Máquina e Maquinaria da Moral”
é da autoria de
Luís Moniz Pereira e António Lopes
e será publicado pela NOVA.FCT Editorial em meados de 2019

OBJECTIVO

Trata-se de um livro de divulgação científica e índole cultural, provisoriamente intitulado “Moral da Máquina e Maquinaria da 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 artigoscientíficos especializados ultimamente publicados por Luís Moniz Pereira, bem como entrevistas e palestras proferidas por este cientista, praticamente na totalidade em Inglês, e os quais serão 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 será 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 articulará três dimensões da questão. Em primeiro lugar,consistirá numa abordagem ao conceito de inteligência e ao modo como ele evoluiu;
em segundo lugar uma abordagem aos tópicos da Economia e sociedade, especulandosobre impactos vários da IA na vida concreta das pessoas. Por fim enfrentar-se-á 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.

A versão final do livro terá a dimensão aproximada de duzentas e cinquenta páginas,assim distribuídas: As primeiras sessenta destinar-se-ão à exploração evolucionária do conceito de inteligência; as setenta seguintes analisarão os variados impactos sociais e económicos evidenciando a necessidade de uma moral social reconfigurada; as cento e vinte páginas finais serão destinadas 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 Evolucionárias, endereçando a moral como um caso da teoria dos jogosevolucionários, produziu um conjunto muito extenso de artigos científiicos 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 instituiçõ es 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 evolucioná ria e aprofundados através de modelos testáveis em cenários artifíciais, como agora permitido pelos computadores. À medida que a investigação avança, podemos conhecer melhor os processos inerentes a 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 a dignidade que lhe é inerente, bem assim como a 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 robô s 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 Histó ria 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**, ***Scienti4ic 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:

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 Arti\icial. (Col. Centro d eDireito 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 willincreasingly and acutely widen the pre-exiting 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 sotfware 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 ins)tutes),

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 soLware 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.