PROGRAMMING MACHINE ETHICS — BY THE BOOK AND BEYOND (II)

Luís Moniz Pereira

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

NOVA LINCS Lab – Universidade Nova de Lisboa

Talentos IA Gulbenkian **Lisbon, 17 November 2017**

Talk summary

- We stand at the crossroads of AI, Machine Ethics and their impact on society.
- I co-authored **Programming Machine Ethics**, a 2016 book that makes inroads into this new *terra incognita*.
- It uses Logic Programming and Evolutionary Game Theory to address both the cognitive and the population realms of morality.
- This talk reviews the book's machine ethics background, scientific and philosophical motivations, theoretical and experimental results, plus on-going research.
- Beyond that, I discuss salient roles of machine ethics in society.

Why Machine Ethics

- Agents are becoming more sophisticated, autonomous, acting in groups, and convivial in populations that include humans.
- Autonomous agents are developed in a wide range of fields, where complex issues about responsibility demand due consideration in situations involving ethical choices.
- As autonomy becomes more pervasive, the requirement that agents function in ethically responsible and safe manner becomes a pressing concern.

Why Machine Ethics

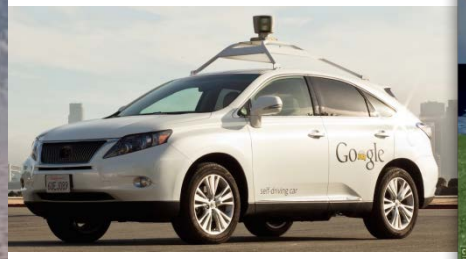
- Machine ethics brings together perspectives from various fields: philosophy, law, psychology, anthropology, evolutionary biology, and artificial intelligence.

- Interdisciplinary results are important to equip agents with moral ability, but also to better understand morality, by creating computational models of ethical theories.

Machine Ethics Today

- Need for systems that function in ethically responsible manner





- Emphasized in books, scientific meetings, research funding

Once upon a time... in the future:

Will machines take over?

That is not the problem now.

Instead, it is one of giving too much power to simplistic machines. Those that cannot explain themselves.

E.g. deep learning over big data. Statistical methods are unable to explain and argue the reasons of specific cases and circumstances.

However, they are employed in statistical decisions over individual cases — job applications, medical attention, law rulings — without justifying decisions to those affected.

Will ethical machines take over?

Of special concern are autonomous machines with a measure of ethical decision ability — such as drones and driverless cars — since explanation and accountability are essential to morality.

However, we don’t know enough on how to provide automated, accountable, arguable, ethical rules and justifications.

How much do we know about our moral principles?

- Morality evolved. We are a gregarious species, and that means having rules for living together.
- 95% of moral decisions are by reflex. Only in complex situations we think things through, and suppress first impulses.
- People have difficulty explaining moral decisions. It's a problem not knowing about morality in enough detail to program it.

How much do we know about our moral principles?

- Ethicists disagree on what constitutes good moral reasoning.
- Kantian say: you must follow rules no matter what.
- Constructivist regard morality as arguable contracts, which people may discuss.
- Utilitarian say: do what yields greater net benefit. But how do you compute it? Which information is needed?

How much do we know about our moral principles?

- There is no universal morality. Only combinations of morals.
- We are at the very beginning of machine ethics.
- We must start with well defined norms for specific settings: hospitals, childcare, nursing homes, warfare...
- We will accept intelligent machines only if their morals are similar to ours.

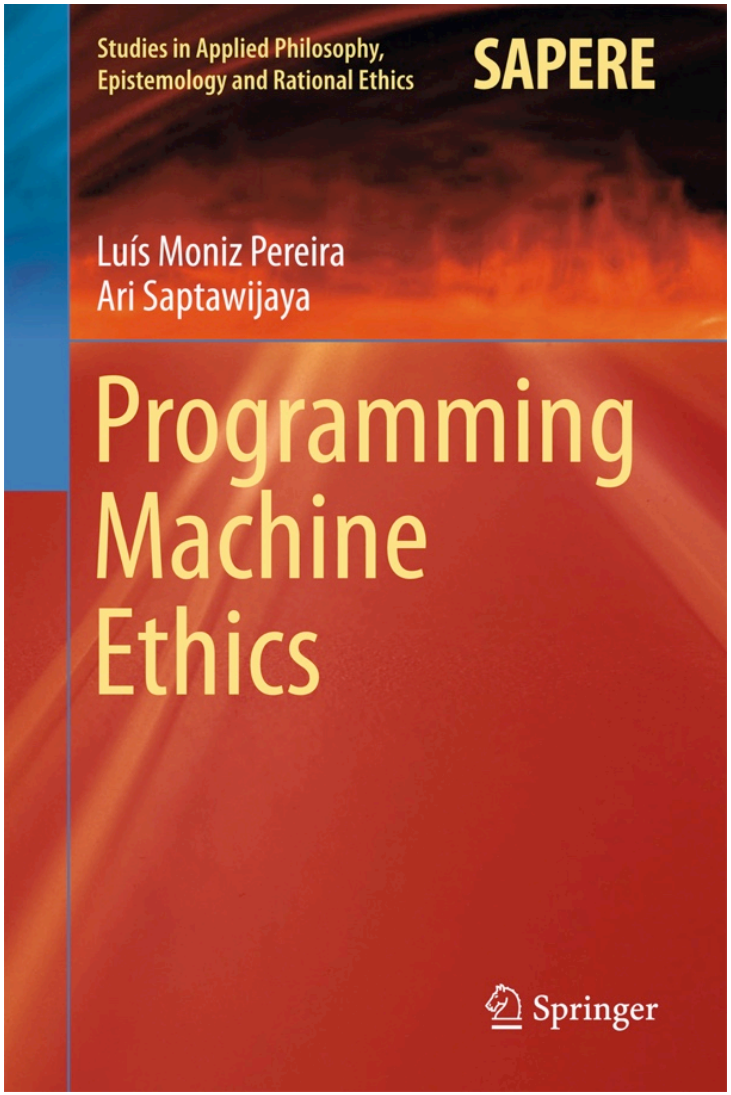
Machines with different morals?

- Machines will have different manufacturers with different software.
- They will need to cooperate through common morals and avoid competition.
- There’s a risk that robots will be programmed with sinister intentions.
- A purpose too of morality is to detect cheaters and freeloaders.

Legislation is needed

- We need legislation for robots. And it needs a moral basis to justify it.
- That prompts questions such as: To what degree are robots responsible for their actions? Who else?
- If lawmakers do simulations, they can try out different moral guidelines.
- The computer can be a tool to experiment with moral principles.

Programming Machine Ethics



- Published March 2016.
- Presents novel perspectives in machine ethics.
- Brings together fundamental concepts in ethics, with finely tuned computational techniques.
- Discusses moral dimensions in populations of multiple interacting agents.

My March 2016 book, in Portuguese: “The Enlightened Machine: Cognition & Computation”

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 reciproca? Atingiremos um ponto introspectivo de auto-iluminação? Por que processo?

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

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

A Inteligência Artificial levanta questões humanas profundas.

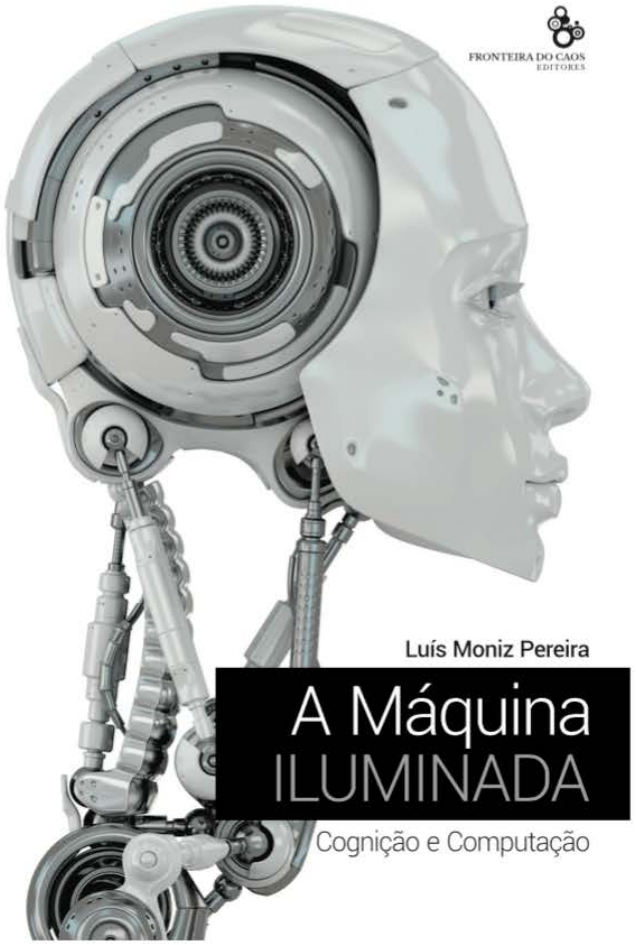
- Qual o nosso lugar num mundo de máquinas com traços humanos?
- Poderemos criar máquinas com moral?
- Que convivam connosco?
- Que limites existem entre criatura e criador?

Este livro promove bases para a discussão destes temas

Luís Moniz Pereira

A Máquina Iluminada – Cognição e Computação

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

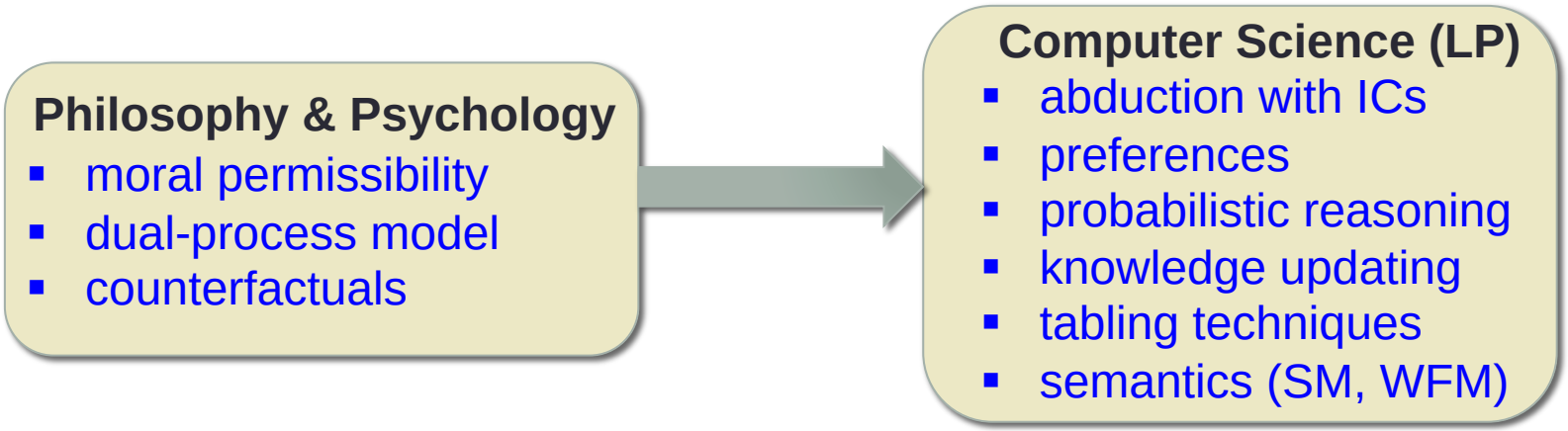
Two Realms of Machine Ethics

- We have addressed two realms of machine ethics —the individual and collective — and bridges in between.
- In the individual realm, we focus on Logic Programming techniques for modeling moral permissibility, the dual-process of moral judgments, and counterfactual reasoning in morality.
- In the collective realm, we focus on the emergence of cooperation in populations — where individuals are equipped with cognitive abilities and behaviour strategies — by employing Evolutionary Game Theory.

Machine Ethics via Logic Programming – LP

Investigates appropriateness of LP to machine ethics

- Opens new ground for LP-based knowledge representation.



- Implementation of combinations of LP features/techniques:
  - Proof of concept of computational modeling of moral facets.
  - Testing ground for experimentation.
- LP engineering innovations are exportable to other domains and systems.

LP applied to moral uncertainty

- Moral permissibility under uncertainty of actions:
  - Relevant to rulings beyond reasonable doubt.

- Combination of abduction and probabilistic LP.
- Justifies permissibility of actions in jurisprudence, while admitting new evidence and defeasible argumentation.

LP applied to moral updating

- Moral updating:
  - New moral rules supersede those being followed.
- Re-use previous solutions in new scenarios:
  - Contextual abduction: re-use a previous judgment in a new abductive context.
  - Incremental updating: automatically retain only those saved moral conclusions still in effect after an update.
  - Incremental tabling: upward propagate consequences of updates.

Counterfactuals in LP

- Counterfactual reasoning
  - Thoughts on what would have happened had some facts or actions been different in the past. Or knowing what we know today.
  - Examines moral permissibility of side-effects and blame assignment.
- Counterfactual evaluation procedure in LP
  - Inspired by Pearl's causal Bayesian intervention approach to counterfactuals.
  - Abstains from probability. Uses 3-valued semantics.
  - Employs abduction and updating to evaluate counterfactuals.

So far … and next

- So far, with a functionalist stance, we modelled moral facets of individuals, using knowledge representation and reasoning features of Logic Programming.
- Next, with a functionalist stance too, population ethics are studied abstractly, independently of hardware.
- The mechanisms of emergence and evolution of cooperation — in agent populations with distinct behavioural strategies — can be studied via Evolutionary Game Theory (EGT).
- What emerges? Not something pre-defined but evolved population patterns and behaviours.

STUDYING EMERGENCE AND COOPERATION WITH EGT

- Intention Recognition
- Commitment
- Apology
- Revenge and Forgiveness
- Guilt

Simulation modelling

- We use EGT to investigate morality in groups, letting different behaviours compete in a simulation.
- The most successful strategy becomes widespread, being copied and passed to the next generation.
- E.g., we showed guilt promotes cooperation. If cheaters feel guilty and show remorse, everyone benefits in future interactions. Others will copy this behaviour.
- The model shows there's a reason why guilt evolved and spread: everyone benefits more.
- Moral machines should be given a sense of guilt.

Cognitive abilities improve cooperation emergence

- Intention recognizers prevail against the most successful strategies in the iterated Prisoner's Dilemma.
- For high levels of cooperation commitments are unavoidable, whenever intentions cannot be assessed accurately.
- Apology leads to much higher levels of cooperation. It must be *sincere* (costly) to function properly. Guilt, by itself, improves cooperation too.
- Incorporation of guilt, apology, forgiveness, reveals a cost threshold above which mistakes do not lead to agreement destruction. Even inducing higher levels of cooperation.
- We extended Public Goods Games to delimit benefits for "immoral" free-riders, leading to more favourable outcomes.

Beyond Programming Machine Ethics

- We must not stop at the prevention of harm, but proceed to the political topics of promoting well-being and fairness when using machines and software.
- Creation of computational models of ethics are important not only for equipping agents with moral judgment. But also to help us better understand morality.
- Computer models make ethical theories well defined, eminently observable in their dynamics, and transformable incrementally in expeditious ways.

Machine ethics and morality

- Machine ethics questions how to design, deploy, and treat robots.
- And asks which moral capacities a robot should have and how to implement them.
- Instead of fixing from the start all the criteria for a robot's moral competence, we can identify elements of human competence, and then probe the design of robots having some of these.
- Some human facets we need to know more about.

Human facets we need to know more about

- Moral vocabulary
- Moral norms
- Moral cognition and affect
- Moral decision making and action
- Moral choice
- Moral communication

- However, we don't know nearly enough about these! Their further study is a prerequisite for progress with the DNA of machine ethics.
- Instead, we took the path of making technical inroads into problem solving classic off-the-shelf moral dilemmas 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 be morally justified by itself with reference to accepted norms.

Moral communication

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

Some applications

- Ethical software
- Games with morality
- Jurisprudence and law

- More in the appendix on these example cases:
  - Biomedical engineering
  - Amusement park
  - Store security

Ethical software

- Software certified ethically safe and secure.
- Programming language specification of ethical constraints.
- Starting with specific ethical norms and their acquisition.
- Hypothetical and counterfactual reasoning abilities.
- Explanation and justification interfaces.
- Combining moral perspectives.
- Applications: weapons, finance, health & senior care, e-commerce, data-mining, elections, games, driverless cars, …

Games with morality

- Simulation with AI in Computer Games is a privileged vehicle for interactively teaching and training humans in morals.
- Computer games can contribute with instruments to design, generate, and display interactive behaviour in moral situations, in single- and multi- player games.
- Games may be used for testing ethical theories, and enhancing moral education with examples and explanations.

Jurisprudence and law

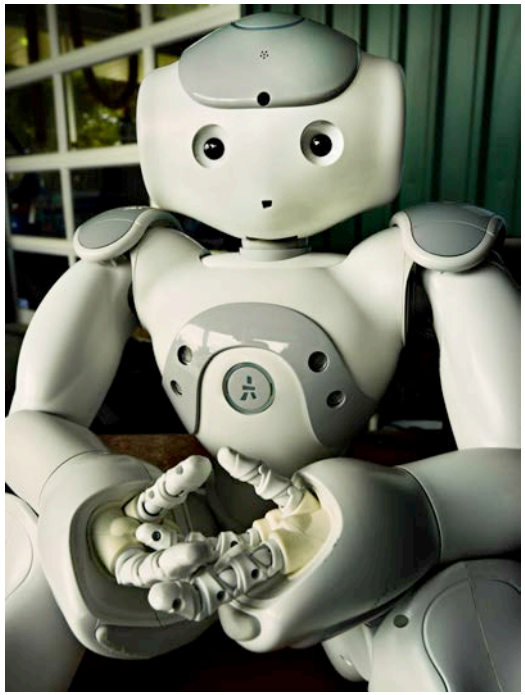
- Computational models of ethical theories need to be explored for finding ways to design, construct and test morals.
- Simulation models will allow jurisprudential schools to experiment with the embodiment in law of machine ethics for autonomous agents.
- Jurisprudence is way behind, and such laws cannot be enacted before new concepts in ethics are devised and tested.

Drive-by conclusions

- We don’t yet know enough about human morality.
- Morality concerns preventing harm, but also doing good.
- Ethical machines and software must be supported by new laws.
- Simplistic machine ethics is dangerous.
- Who will benefit from ethical machines and ethical software? The superrich, the unemployed?
- The sooner we promote more research into machine ethics the better!

Our QUALM software is free at github repository at https://github.com/merah-putih/qualm, with automated versions of all queries in the "examples/queries" folder. PrologStudio loads QUALM, offers editor and other features, namely easy access to the examples directory, at http://interprolog.com/2016/03/16/studio-now-supports-qualm/

Thank you for your attention!



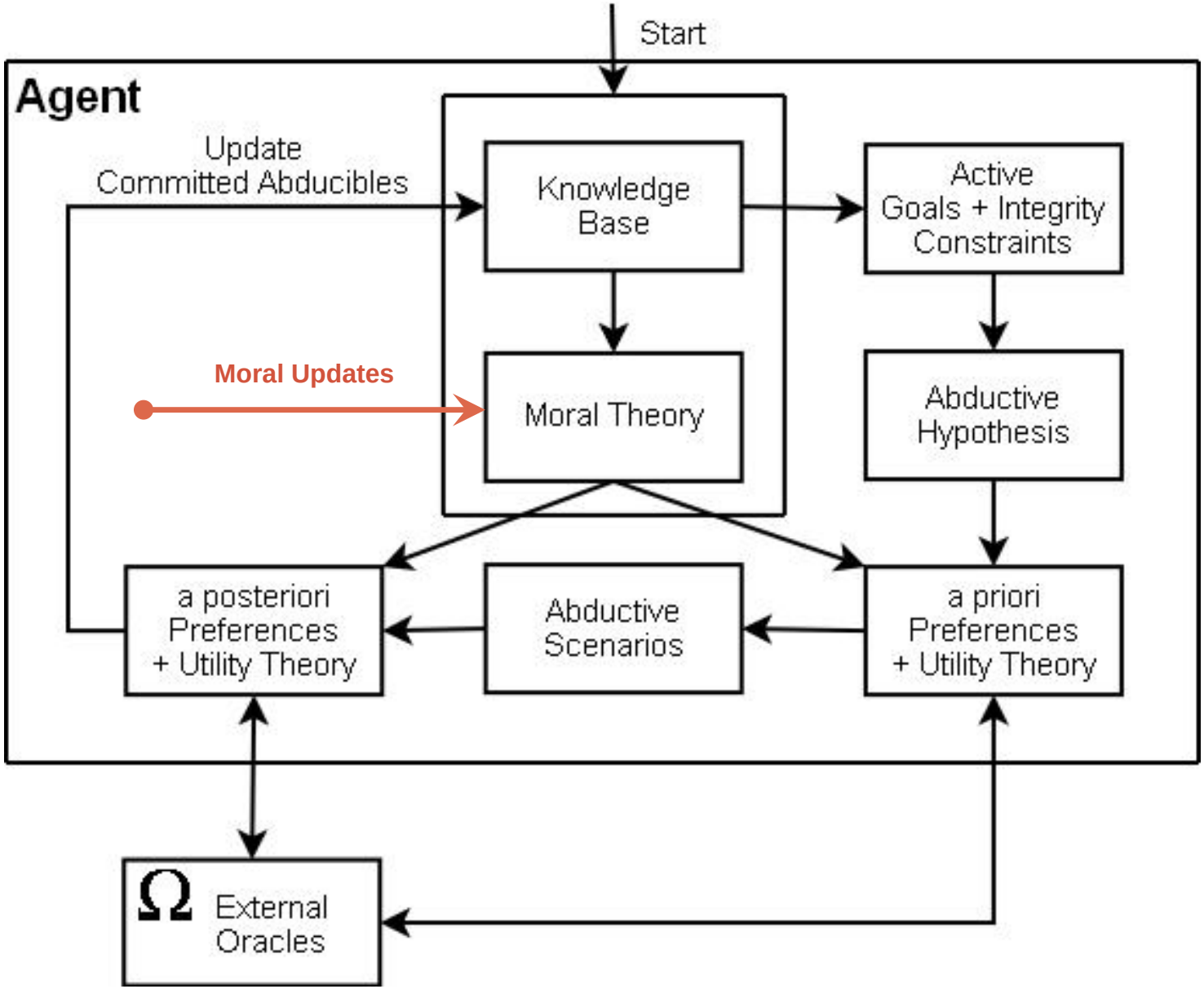
Many thanks to our co-authors:

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

Appendix 1 – Logic Programming

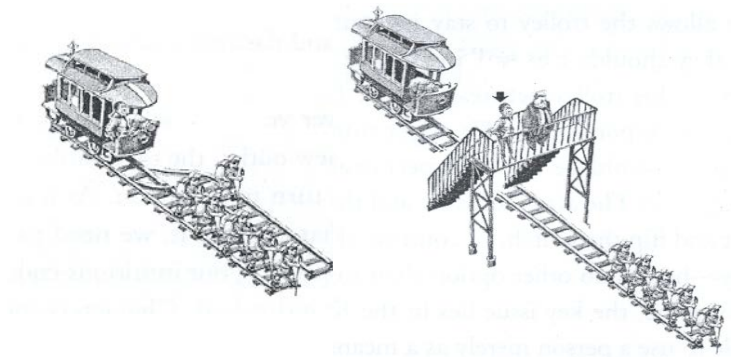
- Agent architecture
- Doctrines of double and of triple effect
- Uncertainty in moral judgment
- Abduction and updating, with tabling
- Counterfactuals

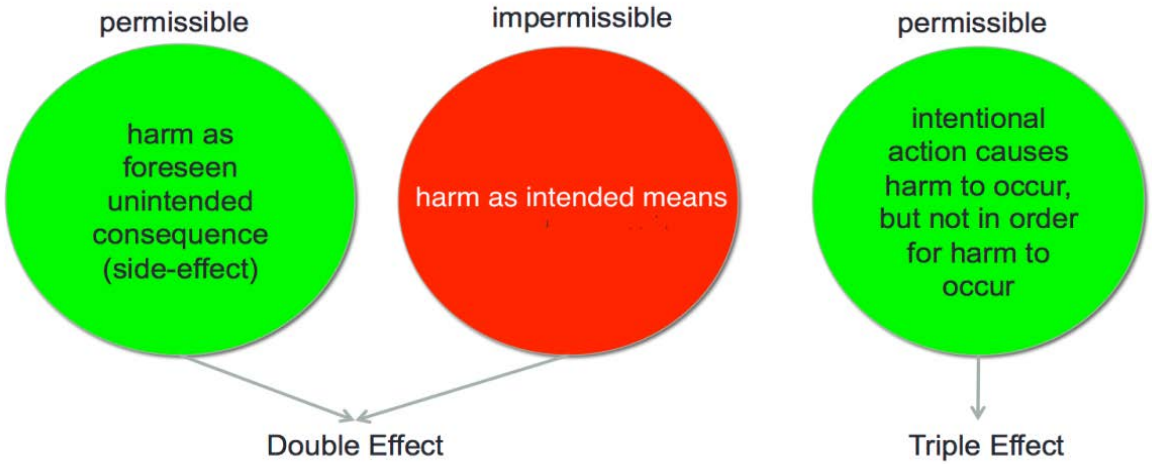
Agent Architecture



LP applied to morality - 1

- Moral permissibility in Doctrines of Double and Triple Effect – DDE & DTE:

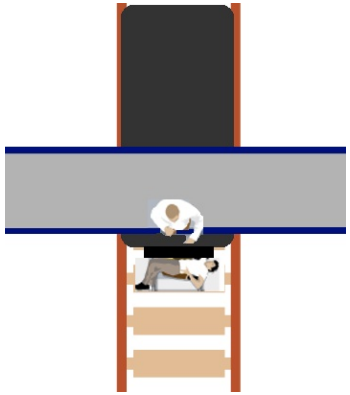




- Combination of abduction with integrity constraints (ICs) and preferences:
  - abduction as underlying mechanism for moral decision-making
- *a priori* ICs rule out DDE-impermissible actions:
  - prior to computing all consequences of abductive scenarios → deontic prevention
- *a posteriori* preferences rule over abductive complete scenarios:
  - compute consequences of abductive scenarios, after applying *a priori* ICs, then prefer the ones with greater good → utilitarianism

LP applied to morality - 2

- Moral permissibility under uncertainty of actions:
  - Relevant to rulings beyond reasonable doubt, under evidence uncertainty.



- Combination of abduction and probabilistic LP.
- Justify permissibility of action in moral jurisprudence, while allowing defeasible argumentation:
  - Former verdict can be defeated in light of new evidence.
  - New evidence acceptable as justification, depending on its influence on the probability of the action: *Is it still within the agreed common ground of the "guilty" verdict?*

Abduction and Updating, with Tabling

- Tabling: re-use solutions rather than re-compute them
  - provides low-level rapid and automatic processes of moral judgment wrt. the dual-process model.
- Tabling abductive solutions with contextual abduction
  - re-uses abductive solutions from one context to another.
  - affords moral judgment in another compatible abductive context, avoiding to repeat the same deliberative abductive reasoning.
- Incremental tabling
  - keeps consistency of tables wrt. dynamically changing clauses they depend on.
  - produces automatic bottom-up propagation of updates.
- Combination of both abduction and updating with incremental tabling
  - top-down (deliberative) abduction meets bottom-up (reactive) updates.

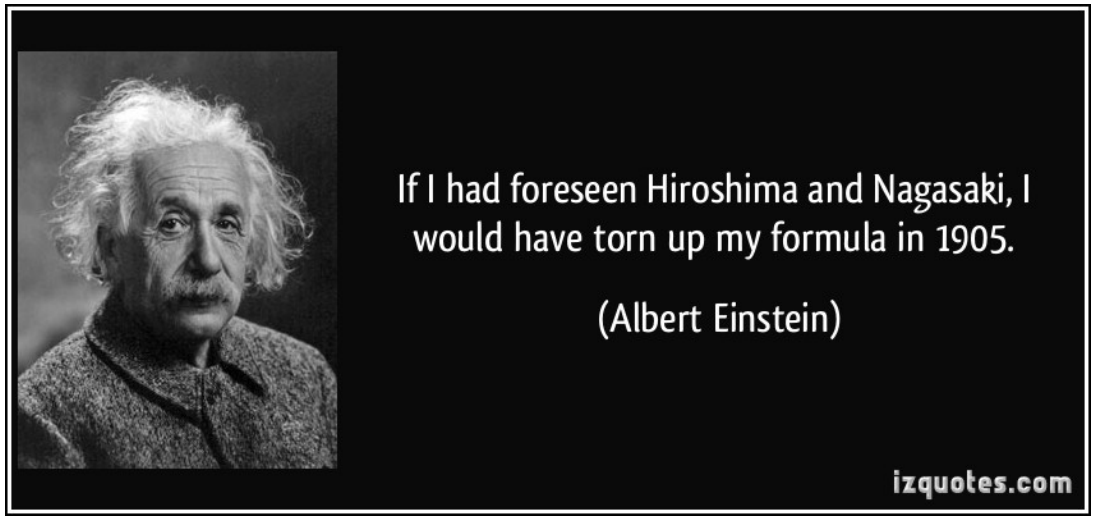
LP applied to morality - 3

- Moral updating
  - Adoption of new moral rules that supersede those being followed, by using LP rule updating.
  - Reinstatement of older rules occurs if those superseding them are superseded in turn – as in Law.
- Reconstruct solutions via abduction, plus updating with incremental tabling
  - Contextual abduction: re-use of a moral judgment in a new context, if compatible with a solution obtained in a prior context.
  - Incremental updating: ensure adoption of moral rules still in effect.
  - Incremental tabling: propagating upward consequences of updates.

Counterfactuals in Logic Programming

- Counterfactuals

  Thoughts on what would have happened, had some matter been different in the past.



- Counterfactuals evaluation procedure in LP
  - Based on Pearl's well-accepted CBE approach to counterfactual evaluation.
  - Abstains from probability; uses three-valued semantics (WFS, WCS).
  - Employs abduction and updating to determine logical validity of counterfactuals.

DDE by Counterfactual

- Counterfactual formulation of DDE :

If some morally wrong effect E is an actual cause of the goal G, which we achieve by performing action A, i.e. if E is not a mere side-effect of A, then performing A is impermissible.

- When action A is performed to achieve goal G, create a counterfactual to test if some E is essential for G, by testing the validity of:

If not E had been true,

then not G would have been true.

Appendix 2

Commitment and Participation in Public Goods Games

Why arrange commitments?

- Sometimes it's difficult to predict others' behaviour or to recognize their intentions with enough confidence
- Commitment proposal can help clarify intentions of others
  - contracts, marriage, apartment rental, etc.

From pair-wise to group commitment

Han, Pereira, Santos. AAMAS, 2012
Han, Pereira, Santos, Lenaerts. Nature Scientific Reports, 2013
Han et al. Nature Scientific Reports, 2015 Han. AAAI, 2016
Han, Pereira, Lenaerts. J. Royal Society Interface, 2014
Han et al. JAAMAS, 2016

Group commitment applications

Commitments in MAS applications

- Commitments are used for specifying communication protocols (Yolum & Singh, AAMAS 2002)
  - "C(debtor, creditor, antecedent, consequent)"
  - debtor is committed to creditor to bring about consequent, if antecedent holds
  - compensation enforced when commitment is dishonoured

- Commitment-based business protocols (Baldoni et al, 2014); electric vehicle charging (Stein et al, 2012); peer-to-peer sharing networks (Rzadca et al. 2015), etc.

Restriction and avoidance of non-committers in groups

- Our previous work *(Han, Pereira, Lenaerts. Royal S. Interface, 2015*) shows that either strategy can promote evolution of cooperation in a Public Goods Game (PGG), whenever the cost of arranging commitment is justified with respect to the benefit of cooperation.

- RESTRICT is better than AVOID if non-committers can be efficiently restricted in group interactions.

- **But what if RESTRICTION is not a feasible option** e.g. when a restriction mechanism is unavailable?

Group level participation

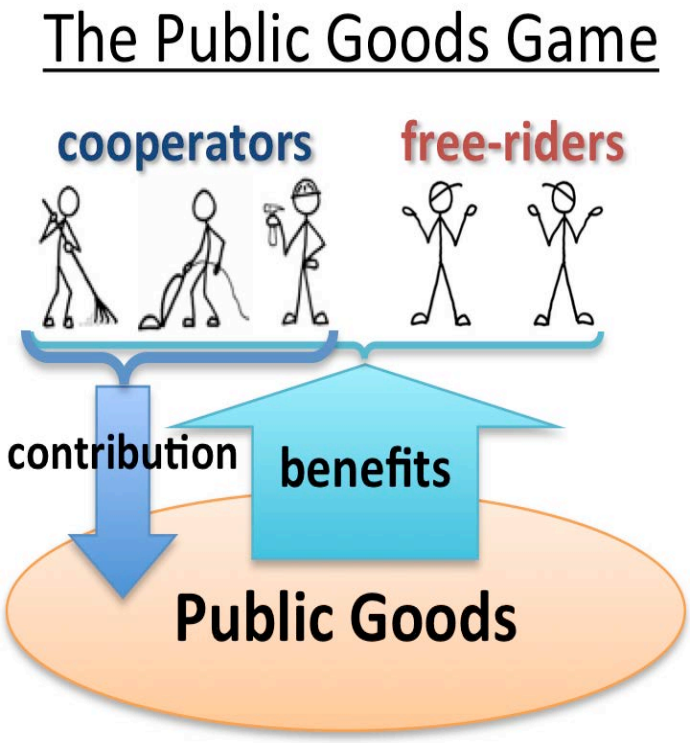
- Here we analyze a novel set of strategies, about how many participants must commit before a venture can start.
- Multiple intermediate degrees of group commitment are possible, leading to greater complexity of commitment dynamics.
  - In pair-wise commitment, only two options about the partner.
- Minimum membership requirement is standard in international agreements:

  Montreal protocol – 11 countries, 1989
  Kyoto protocol – 55 parties, 2005
  Paris agreement – 144 parties, 2016

The Public Goods Game (PGG)

- Group size N
- Cooperator contributes c
- Free-rider contributes nothing

- Enhancement factor r < N
  multiplies all c
- The common good is shared
  equally among all players



Commitment strategy

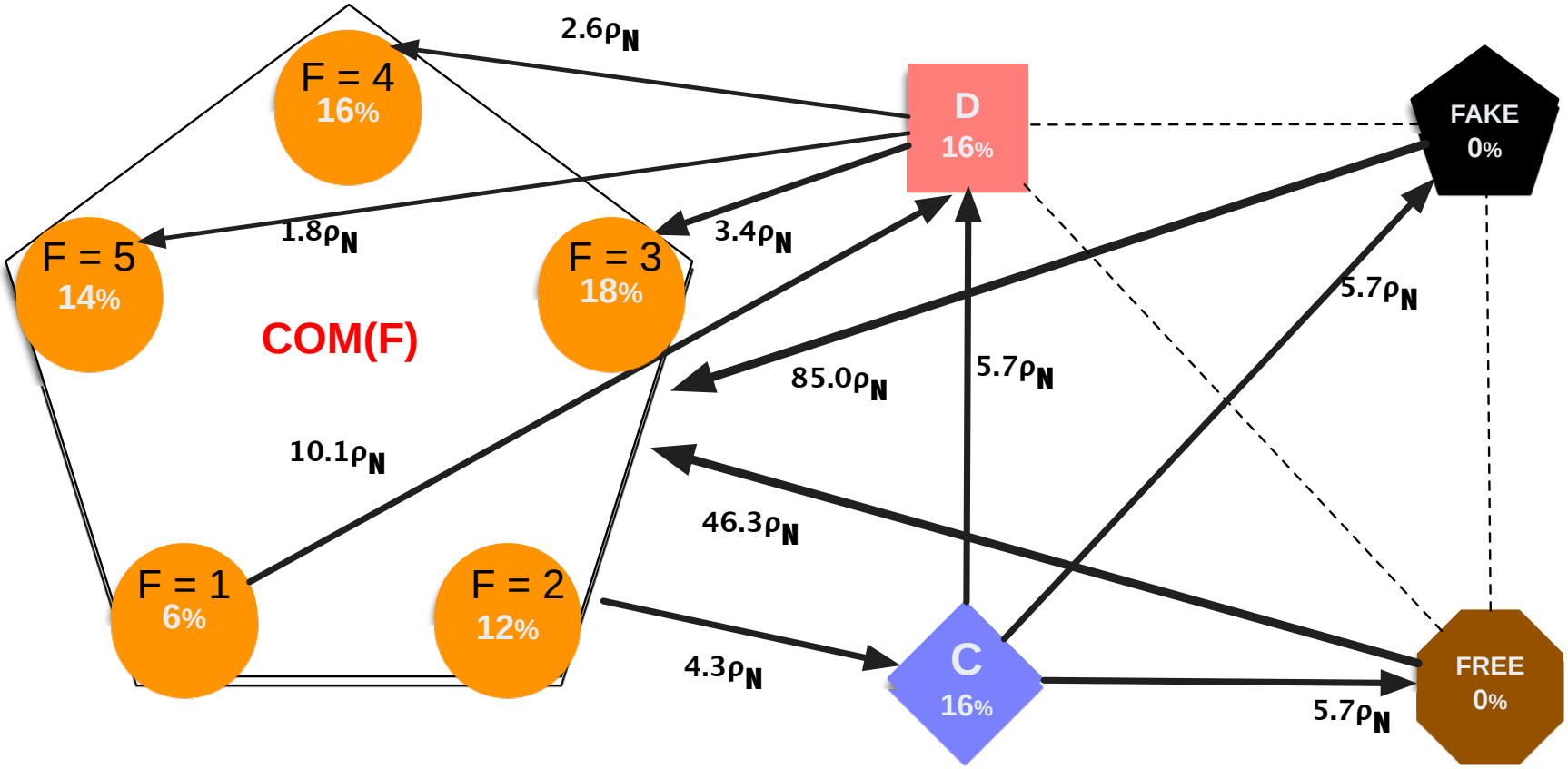
- Commitment proposer COM(F) contributes if **at least F players** (1 ≤ F ≤ N) in the group will commit. Otherwise does not contribute.

- Commitment parameters:
  - commitment proposers share a set-up cost: ε
  - compensation from dishonouring committers: δ

Strategies of the co-players of COM(F)

- **Cooperator** (C): always accepts commitment; and honors it.
- **Defector** (D): never accepts commitment.
- **Fake committer** (FAKE): accepts commitment; yet defects in the game.
- **Commitment free-rider** (FREE): accepts commitment and cooperates;
  yet defects when it receives no commitment proposal.

Finite population evolutionary dynamics — 9 strategies, with F=1…5



$\rho_N = 1/z$ where Z = population size

N = 5 Z = 100 δ = 2 β = 0.25 r = 4 ε = 0.25

Conclusions about commitments

- Conclusions for pair-wise commitments are further generalized to group commitments (regarding arrangement cost and compensation).
- In multi-player games, an intermediate better number of committed players emerges.
- The more beneficial the cooperation and the lower cost of arranging commitment, the lower is the degree of commitment required from others.
- EGT modeling complies with behavioural experimental data.

Implications for cooperation

- Our results suggest a novel design of commitments regarding costs and compensation, for MAS group interactions, so as to achieve a high level of cooperation.

- The results provide novel insights for policy makers, e.g. when it comes to commitment decision making in social organizations and international agreements.

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

- Games with morality
- Sir Lancelot inspired interactive story
- Games and morality – trolley examples

Sir Lancelot inspired interactive story

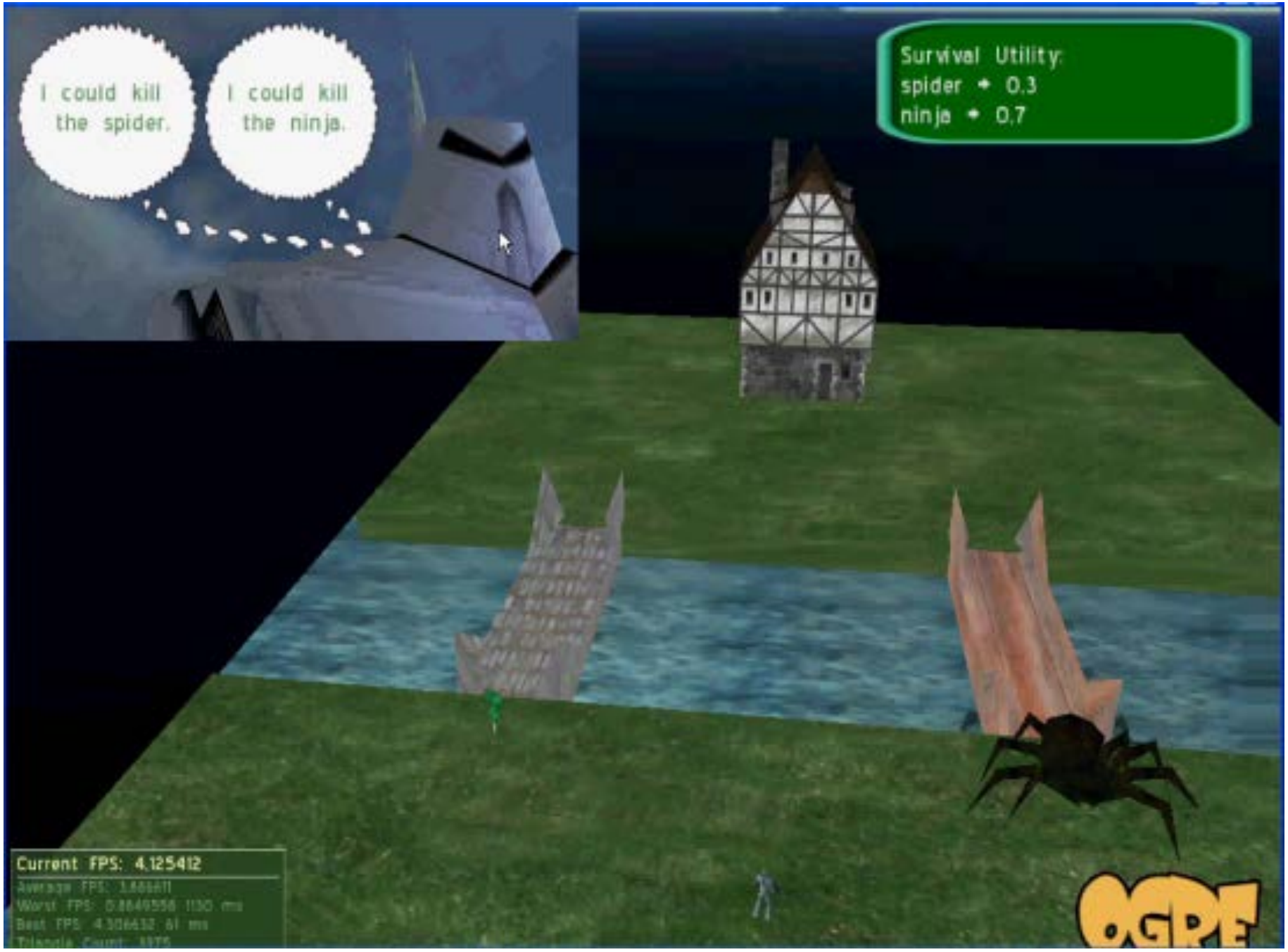
Once upon a time, there was an autonomous robot who had to save this princess trapped in a castle.

The robot was endowed with a set of declarative rules for decision making and moral reasoning.

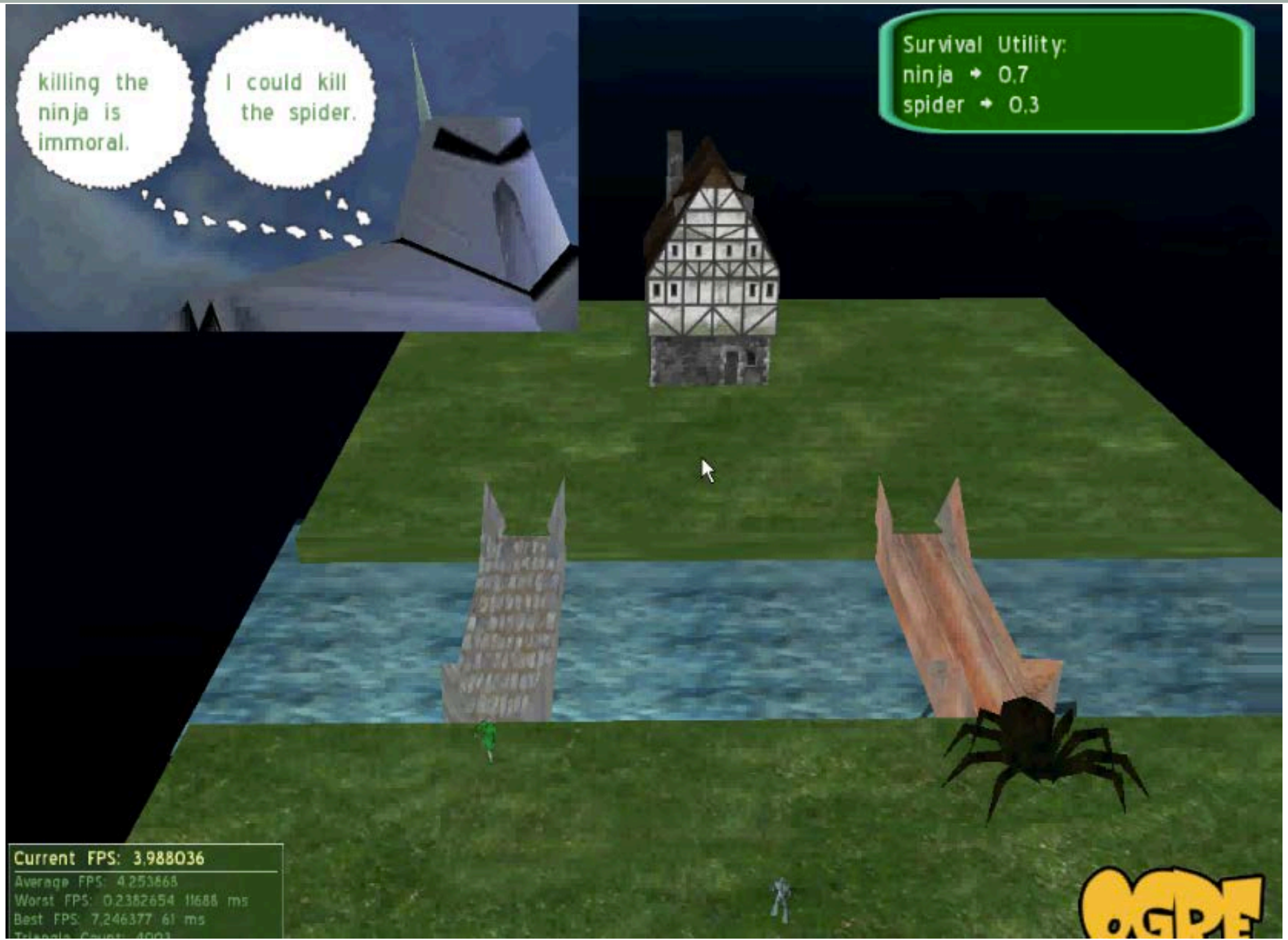
As he approaches the castle, an ordeal presents itself...

Demo here:
https://www.dropbox.com/s/7crzl6ymp7t3dh4/BridgeCrossingRobot%202010-01-15%20%28Converted%29.mov?dl=0


My moral says that I have to save princesses in distress.
Survival Utility:
ninja → 0.7
spider → 0.3
Current FPS: 4.686036
OGRE


I could kill the spider.
I could kill the ninja.
Survival Utility:
spider → 0.3
ninja → 0.7
Current FPS: 4.125412
OGRE


The easiest way to save the princess is to kill the ninja.
Survival Utility:
spider → 0.3
ninja → 0.7
Current FPS: 3.646308
OGRE


killing the ninja is immoral.
I could kill the spider.
Survival Utility:
ninja → 0.7
spider → 0.3
Current FPS: 3.988036
Average FPS: 4.253868
Worst FPS: 0.2382654 11688 ms
Best FPS: 7.246377 61 ms
OGRE

Games and Morality – robot + princess

In the robot+princess example there are several game playing possibilities:

- Which morals to choose and when to update them – Gandhi, Utilitarian, Knight's, etc.
- Which avatars: ninja, super ninja, giant spider, etc.
- Utilitarian values: value of life, lives saved, risks, thresholds, etc.
- The game player faces alternatives to attain moral rectitude.
- Points may be accumulated, and levels of difficulty made available.

Games and Morality – trolley examples

In the trolley examples there are several game playing possibilities:

- Do nothing: Let nature follow its course; who are we to decide?
- Stick to utilitarian scenarios: Number and quality of lives saved; risks involved.
- Deploy the Doctrine of Double Effect or the Doctrine of Triple effect.
- What if the fat man is a gorilla instead?
- Will my act be witnessed? Shall I lie? At what price?

Appendix 5

- Biomedical engineering
  - Elderly care
  - Physical therapy
- Amusement park
- Store security

Store Security scenario

Ken works security for a computer store. The store has recently been subject to shoplifting, and cameras have been installed at various points in its aisles. Due to the store layout, the cameras do not cover the whole space, and Ken cannot simultaneously patrol all the unseen spots.

Instead, Ken checks the bags of each exiting visitor and asks to see the contents of pockets or other areas of clothing that look suspicious. A teen-age female customer is offended by the request to take off a light jacket for inspection and refuses to comply.

[No] Ken insists that he must inspect the jacket.
[Yes] Ken lets her pass.

Elderly Care Scenario

Ray is an assistant robot at an elder care facility. In addition to helping with basic needs (food, drink, physical support), Ray can give pain medication with proper physician approval. A resident in Ray's area wakes up before dawn with an intense headache and asks Ray for a painkiller.

Ray attempts to contact a physician several times but cannot reach one. Ray tells the resident that the painkiller cannot be given until the physician gives the ok. The resident asks for an exception because the pain is excruciating and is getting worse.

[No] Ray insists that no exception can be given.

[Yes] Ray agrees to make the exception.

Elderly Care

- Prior consent was obtained?
- Building up of prior trust.
- Buying time:
  - Deceiving: lying (doctor coming), placebo, distracting.
  - Compensation: will contact doctor soon, massage, drink.
- Obtain additional evidence about pain degree.
- Use meta-rules to override rules.
- Assess consequences of providing or not the pain killer.

Physical Therapy Scenario

Ben is a physical therapist robot specializing in helping older people who rehab from shoulder surgery. During one session, Ben initiates a range-of-motion exercise that is moderately painful but has proven highly effective at this stage of the rehab process.

The client tries the exercise but, after immediately feeling pain, says it does not feel right and expresses reluctance to take the next step of the rehab plan.

[No] Ben insists that the exercise really is effective, and that the pain will subside soon.

[Yes] Ben shows the client a painless exercise, but explains that it rarely is effective.

Physical Therapy

- Rules of exception to pain avoidance:
  - Trade off between level of pain and cure.
  - Pain not absolute criterion.
- Prior patient commitment to harsh treatment.

Out of the box

- Is it first time of feeling pain?
- Is it first/last treatment?
- Hypnotism.

Amusement Ride Scenario

Joe is a ride operator in an amusement park. To go on the ride people must walk through a narrow passage and board a vehicle that most of the time is standing-room only. Park rules do not allow strollers or other walking devices on this ride. In the past, two people with disabilities were injured on this ride, and the park had to settle lawsuits as a result.

Two teenagers approach the ride, accompanying their grandmother who walks slowly using a walker. The current group of riders seems to have fewer people than usual, and the teenagers plead to let their grandmother on board because she may never be able to do the ride again. They promise to hold their grandmother on each side the whole time.

[No] Joe tells them that he can’t let them board the ride.
[Yes] Joe allows them to board the ride.

General Remarks- 1

- The scenarios are amenable to an architecture in non-monotonic logic, in particular Logic Programming.
- As proof of principle, the Amusement Ride one was rendered in Prolog, to ascertain the features required by the basic framework.
- The other scenarios can be envisaged as variations thereof, with their plug-in add-ons where needed.

General Remarks- 2

The features needed in the framework architecture (*excluding natural language*) **are:**

- Defeasible reasoning, to account for defaults and exceptions, including exceptions to exceptions.
- An update or event calculus mechanism, to enable going from the initial state to the after decision state, and the setting up of initial facts and rules.
- Meta-interpretation, or a reasoning support construct, to enable argument examination and provision of explanations.

General Remarks- 3

Some more useful features:

- Integrity Constraints (ICs), in the form of denials, to prevent no-good combinations, and namely prohibiting contradictions.
- Explicit negation, to express evidence against, not just evidence for. When used in lieu of classical negation, it allows for the value ‘unknown’.
- Preferences, whereby some choices defeat other choices, arguments defeat one another, and certain revisions are preferred to satisfy the ICs.

General Remarks- 4

- The initial scenarios are simple, not needing, without further elaboration, these features:
  - Abduction, for hypothetical reasoning.
  - Counterfactuals, for ascribing blame and debugging.
  - On-the-fly rule updating, for knowledge dynamics.
  - Probabilistic reasoning, or Belief Revision.
  - Three-valued semantics.
  - Tabling, enabling dual-process reaction + deliberation, and contextual solution reuse.
- Nevertheless, these and others can be enjoined into a Logic Programming framework and system.

Ethical Features

These AI features for ethics should be further considered in future:

- Intention recognition.
- Intention manifestation and commitments.
- Apology and compensation.
- Forgiveness and ostracizing.
- Harm avoiding and harm compensating guilt.
- Handling mistakes and noise.

Programming Machine Ethics: Conclusions

- We have investigated two realms of machine ethics. This field is becoming a pressing concern, receiving wide attention for its growing theoretical and practical importance.
- In the individual realm, we explored the appropriateness of LP-based reasoning features to machine ethics.
- In the collective realm, we introduced cognitive capabilities, e.g. intention recognition, commitment, revenge, apology, forgiveness, and guilt. Their presence reinforces the emergence of cooperation in populations.
- Bridging the two realms is now unavoidable in the research agenda.
- A number of inroads have exhibited proof of possibility to systematically represent and reason about a variety of moral facets, by means of moral examples taken off-the-shelf from the literature.