Evolutionary Game Theory Modelling of Guilt

Luís Moniz Pereira [1] and **Tom Lenaerts** [2] and **Luis A. Martinez-Vaquero** [3] and **The Anh Han** [4]

Abstract. Inspired by psychological and evolutionary studies, we present two theoretical models wherein agents have the potential to express guilt, with the ambition to study the role of this emotion in the promotion of pro-social behaviour. We show that the inclusion of the emotion of guilt, in the sense arising from actual harm done to others from inappropriate action or inaction, is worthwhile to incorporate in evolutionary game theory models of cooperation, for it can increase cooperation by correcting and inhibiting defection. The abstract study thereof profitably transpires to concrete considerations in the design of artificial multi-agent populations. To achieve this goal, analytical and numerical methods from evolutionary game theory have been employed, but not shown in too fine detail here, to identify that reasonable conditions exist for which enhanced cooperation emerges within the context of the iterated prisoners dilemma. Guilt is modelled explicitly as two features, i.e. a counter that keeps track of the number of transgressions and a threshold that dictates when alleviation (through for instance apology and self-punishment) is required for an emotional agent. Such alleviation introduces an effect on the payoff of the agent experiencing guilt. We show that when the system consists of agents that resolve their own guilt without considering the co-player's attitude towards guilt alleviation then cooperation does not emerge. In that case, agents expressing no guilt or having no incentive to alleviate the guilt they experience easily dominate the guilt prone ones. On the other hand, when the guilt prone focal agent requires that guilt only needs to be alleviated when guilt alleviation is also manifested by a defecting co-player, then cooperation may thrive. This observation proves consistent in a generalised model discussed in this article. In summary, our analysis provides important insights into the design of multi-agent and cognitive agent systems, wherein the inclusion of guilt modelling can improve agents' cooperative behaviour and overall benefit.

1 INTRODUCTION

> "...what do you think, if a person does something very bad, do they have to be punished?"..."You know the reason I think they should be punished?"..."It's because of how bad they are going to feel, in themselves. Even if nobody did see them and nobody ever knew. If you do something very bad and you are not punished you feel worse, and feel far worse, than if you are." [Page 55 of "The love of a Good Woman" by Alice Munro (Nobel Prize in Literature 2013) in "Family Furnishings-Selected Stories 1995-2014", Vintage Intl. Edition, 2015]

Presently there is a general mounting interest on machine ethics [22] and recent research monographs have been addressing its issues [17]. One concerns the computational modelling of human emotions, amongst which we find guilt and its role in minimising social conflicts [14]. Guilt is defined in the online Merriam-Webster dictionary as "The feeling of culpability especially for imagined offences or a sense of inadequacy", which implies that guilt follows from introspection: An individual experiencing guilt will detect this emotional state, and can act upon it. Guilt is an evolved pervasive feature in human cultures, which can lead to enhanced cooperation via changes in behaviour or upon apology (cf. background references below). Frank argued that guilt may provide a useful mechanism, if operationalised properly, to miminise social conflict and promote cooperation [3]. Notwithstanding the importance of this emotion for the evolution of cooperation, no in-depth numerical or analytical models have been provided to confirm or refute the hypothesis that this emotion has evolved to ensure stable social relationships. Hence, it is natural to enquire how it might enhance cooperation in evolving artificial multi-agent systems, by means of machine implemented models of guilt. With that in mind, we avail ourselves of Evolutionary Game Theory (EGT) [12, 21] to conclude that under certain conditions cooperation can be enhanced by a *modicum* of guilt in a population of autonomous agents.

A distinct evolutionary and population sensitive EGT model of guilt has been explored in [20]. They focus on behaviours associated with guilt, such as apology, but do not however explicitly represent any self fitness changes from the experience of the guilt emotion, like we do in our models. Moreover, their guilt prone agents (GP) do not initiate defection like ours do, but defect only in reaction to another's defection, though they will then feel guilty for having done so. Instead, we crucially associate guilt with self-punishment, and show how this affecting of fitness can be conducive to a population beneficial Evolutionary Stable Strategy (ESS) state [21], one towards which the population evolves to play the strategy, and which state cannot be invaded by a small number of agents using a different strategy. This is the case in our improved (second) model, where self-punishment is only enacted if the other party is not recognised to be guilty too. In [4] (non-evolutionary) utilitarian game theory is employed to model the behaviour resulting from guilt, not by introducing self-punishment but by introducing a guilt aversion level term into a player's utility function, which takes into account the agent's history of previous pairwise interactions and individually learning from it. In contrast, our moral stance to guilt is not utilitarian, in the sense that no individual measure of greater good is being explicitly optimised. We rely instead on social learning in a population's emergent evolution, without recourse to individual histories. Hence our approach and results are thus distinct from previous ones in important ways. Next we frame our hypotheses on guilt within EGT and define our models and methods. Thence we proceed to the presentation of results, and wrap

[1] Universidade Nova de Lisboa, Portugal, email: lmp@fct.unl.pt
[2] MLG, Université Libre de Bruxelles and AI lab,Vrije Universiteit Brussel, Belgium, email: tom.lenaerts@ulb.ac.be
[3] Institute of Cognitive Sciences and Technologies, Rome, Italy, email: fnxabraxas@gmail.com
[4] Teesside University, UK, email: T.Han@tees.ac.uk

up with some justified conclusions and future work.

2 EVOLUTIONARY GAME THEORY MODEL FOR GUILT

Considering the foregoing, an attempt to introduce guilt in EGT models of cooperation seems unavoidable. The issue concerning guilt within such models is whether its presence is more worthwhile than its absence, with respect to a possibly advantageous emergence of cooperation. One can introduce guilt explicitly in models to show that it is worthwhile, in further support of its appearance on the evolutionary scene. Indeed, one may focus on emotions, like guilt, as being strategies in abstract evolutionary population games, sans specific embodiment nor subjective *quale* [18].

We can test this hypothesis via one model spelled out below, whose details can be found in [16]. In it guilt is tied to intention recognition, since it will have evolved as a fear about the detection of harm done (see above). The prediction is that guilt will facilitate and speed-up the emergence of cooperation. In spite of its initial heavier cost, in time that cost will be recuperated within the guilt-ridden population, via inhibition of defection as a result of guilt avoidance. Furthermore, one's timely recognition of another's prior give away guilt signs, on account of her actual intent to harm, can prevent one's self-punishing guilt in cases it would be uncalled for. The base hypothesis is thus that when there exists guilt in the starting population then the most frequent stationary distribution includes the incorporation of guilt and enhances overall cooperation. For which parameters of guilt this happens can be analytically determined experimentally.

2.1 Models and methods

A behavioural quantification of guilt provides us with a basis to define our evolving agents: Guilt is part of an agent's representation or *genotype*, i.e. they will all be equipped with a guilt threshold G, with $G \in [0, +\infty]$, and a transient guilt level, g ($g \geq 0$). Initially g is set to 0 for every agent. If an agent feels guilty after an action that she considers as wrong, then the agent's g is increased (by 1). When g reaches the agent's guilt threshold, i.e. $g \geq G$, the agent can (or not) act to alleviate her current guilt level. We assume here that guilt alleviation can be achieved through a sincere apology to the co-player or, otherwise, through self-punishment if it is not possible to apologise [1, 6]. Different from prior work [8, 15], we do not assume here that apology leads to a benefit for the co-player, considering it only as an honest signal of the experiencing of guilt. In general, the cost of guilt alleviation is modelled by a so-called *guilt cost* γ ($\gamma \geq 0$). Whenever the agent punishes herself, by paying γ, g is decreased (by 1). Using this genotype definition, one can imagine different types of agents with different G thresholds, such as those who never feel guilty (the unemotional ones, with $G = +\infty$) or those who are very emotional, feeling guilty immediately after a wrongdoing (with $G = 0$).

The objective of this work is to show that agents expressing this emotion, despite the disadvantage of the costly guilt-alleviation acts, are evolutionary viable, can dominate agents not expressing the emotion and that they induce sustained social interactions, all of which will be shown in the context of the Iterated Prisoner's Dilemma (IPD). To set the stage for future work we first focus on two extreme behaviours, i.e. $G = 0$ and $G = +\infty$, as will be explained in more detail later. These results are generalisable to situations where $G > 0$ yet less than the number of rounds in the IPD, since when G is larger this would correspond to $G = +\infty$. We use a stochastic evolutionary model incorporating frequency-dependent selection and mutation to identify when agents with guilt are evolutionary stable [21]. More importantly, we will show that for guilt to be evolutionary viable, it should be reactive to the guilt-driven behaviour of the co-player: If the other party is not behaving properly and/or does not show guilt-alleviating behaviour then the focal agent's guilt is alleviated automatically or even non-existing. Pure self-punishment without social considerations will not allow for guilt to evolve at the individual level. In this sense, our work contrasts with for instance that of Gadou et al. [4] which takes an utilitarian perspective to model the behaviour resulting from guilt, not by introducing self-punishment but by introducing a guilt aversion level term into a player's utility function, which ignores the social role of guilt [3]. From a multi-agent perspective, considering socio-technical systems including autonomous agents, our results confirm that decision making conflicts can be reduced when including emotions to guide participants to socially acceptable behaviours.

2.2 Iterated prisoner's dilemma (IPD)

Social interactions are modelled in this article as symmetric two-player games defined by the payoff matrix

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} R, R & S, T \\ T, S & P, P \end{pmatrix} \end{array}$$

A player who chooses to cooperate (C) with someone who defects (D) receives the sucker's payoff S, whereas the defecting player gains the temptation to defect, T. Mutual cooperation (resp., defection) yields the reward R (resp., punishment P) for both players. Depending on the ordering of these four payoffs, different social dilemmas arise [12, 21]. Namely, in this work we are concerned with the PD, where $T > R > P > S$. In a single round, it is always best to defect, because less risky, but cooperation may be rewarding if the game is repeated. In IPD, it is also required that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection ($2R > T + S$); otherwise alternating between cooperation and defection would lead to a higher payoff than mutual cooperation. The PD is repeated for a number of rounds, where the number of rounds is modelled by Ω.

2.3 Guilt modelling in IPD

Starting from the definition of the agent-based guilt feature in the Introduction, we will focus in the current work only on two basic types of (extreme) guilt thresholds:

- $G = +\infty$: In this type of agents the guilt level g will never reach the threshold no mater how many times they defect; hence, they never need to reduce g, and consequently never pay the guilt cost γ. Experiencing no guilt feeling, these agents are dubbed (guilt-) unemotional.
- $G = 0$: whenever this type of agents defects, it becomes true that $g > G$; hence, the agents need to act immediately to reduce g, thus paying γ. These agents always feel guilty after a wrongdoing, viz. defection, and are dubbed (guilt-) emotional agents.

Besides the guilt threshold, an agent's strategy is described by what she plays in a PD (C or D) and, when the agent's ongoing guilt level g reaches the threshold G, by whether the agent changes her behaviour from D to C. Hence, there are five possible strategies, thus labeled:

1. Unemotional cooperator (C): always cooperates, unemotional (i.e. $G = +\infty$)
2. Unemotional defector (D): always defects, unemotional (i.e. $G = +\infty$)
3. Emotional cooperator (CGC): always cooperates, emotional (i.e. $G = 0$)
4. Emotional non-adaptive defector (DGD): always defects, feels guilty after a wrongdoing (i.e. $G = 0$), but keeps behaviour.
5. Emotional adaptive defector (DGC): defects initially, feels guilty after a wrongdoing (i.e. $G = 0$), and behaviour goes from D to C.

In order to understand when guilt can emerge and promote cooperation, our EGT modelling study below analyses whether and when emotional strategies, i.e. those with $G = 0$, can actually overcome the disadvantage of the incurred costs or fitness reduction associated with the guilt feeling and its alleviation, and in consequence disseminate throughout the population. Namely, in the following we aim to show that, in order to evolve, guilt alleviation through self-punishment can only be evolutionarily viable when only the focal agent misbehaves. In other words, an emotional guilt-based response only makes sense when the other is not attempting to harm you too. To that aim, we analyse two different models, which differ in the way guilt influences the preferences of the focal agents, where the preferences are determined by the payoffs in the matrices (1) and (2).

In the first model, an agent's ongoing guilt level g increases whenever the agent defects, regardless of what the co-player does. The payoff matrix for the five strategies C, D, CGC, DGD, and DGC, can be written as follows

$$\begin{array}{c|ccccc} & C & D & CGC & DGD & DGC \\ \hline C & R & S & R & S & \frac{S+R\Theta}{\Omega} \\ D & T & P & T & P & \frac{P+T\Theta}{\Omega} \\ CGC & R & S & R & S & \frac{S+R\Theta}{\Omega} \\ DGD & T-\gamma & P-\gamma & T-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\ DGC & \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S(\Theta)}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} \end{array}, \quad (1)$$

where we use $\Theta = \Omega - 1$ just for the purpose of a neater representation. Note that the actions C and CGC are essentially equivalent; both considered for the sake of completeness of the strategies set.

The entries in the matrix are derived as follows. For instance, when a C player interacts with another C (resp. D) player, it always obtains payoff R (resp. S), in all the rounds of the IPD, so it obtains the same payoff on average, as indicated in the payoff matrix. When C interacts with DGC, it obtains S in the first round and then R in the remaining $\Omega - 1$ rounds (thus it obtains $\frac{S+R(\Omega-1)}{\Omega}$ on average), as the DGC player feels guilty after defecting in the first round, thereby switching to C. Respectively, DGC obtains T in the first round and then R in the remaining $\Omega - 1$ rounds, i.e. $\frac{T+R(\Omega-1)}{\Omega}$ on average. As in this model DGC does not take into account the co-player's attitude towards guilt alleviation, when interacting with D it defects in the first round then changes to C, even when the co-players shows no sign of guilt feeling.

In the second model, an agent feels guilty when defecting if the co-player acted pro-socially or was observed to feel guilty after defection, viz. through exercising self-punishment or apologising. Thus in this second model, guilt has a particular social aspect that is missing from the first model. In particular, DGC does not change behaviour to C if the co-player played D and did not try to alleviate her guilt as a result of her bad behaviour. Now, the payoff matrix is rewritten:

$$\begin{array}{c|ccccc} & C & D & CGC & DGD & DGC \\ \hline C & R & S & R & S & \frac{S+R\Theta}{\Omega} \\ D & T & P & T & P & P \\ CGC & R & S & R & S & \frac{S+R\Theta}{\Omega} \\ DGD & T-\gamma & P & T-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\ DGC & \frac{T-\gamma+R\Theta}{\Omega} & P & \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} \end{array}. \quad (2)$$

The difference can be seen in the new payoff obtained by DGC when playing with D. It no longer changes from D to C after defecting in the first round, thus obtaining P in all the rounds. Notice the differences in the payoff matrices for the interactions between the emotional strategies that defect, i.e. DGD and DGC, and the unemotional defector D.

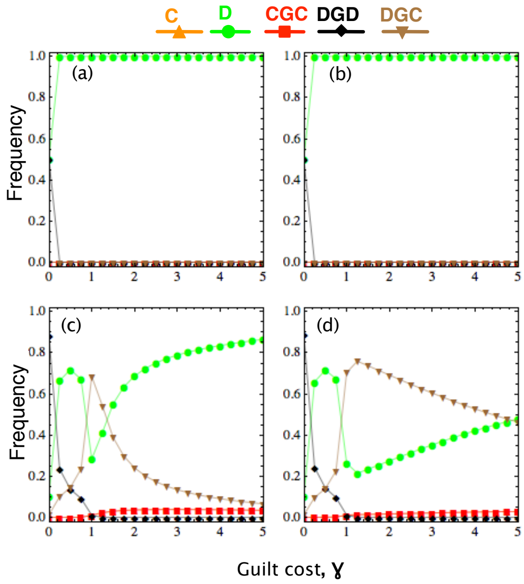

Figure 1. Frequency of each strategy as a function of the guilt cost, γ, for the two models, and for different PD game configurations (see below). In the first model (panels a and b), D always dominates the population. In the second model (panels c and d), for an intermediate value of γ, DGC is the most frequent strategy; but when it is too small or too large, DGD is dominant. Parameters: $\beta = 1$; $N = 100$; $\Omega = 10$; In panels (a) and (c): $T = 2,\ R = 1,\ P = 0,\ S = -1$; In panels (b) and (d): $T = 4,\ 3 = 1,\ P = 0,\ S = -1$.

2.4 Results

We have elsewhere [16] derived analytical conditions (not proffered here) for when DGC can be a viable strategy, which is risk-dominant when playing against defection strategies (i.e. D and DGD). We have shown that though the DGC strategy is always dominated by defective strategies in the first model, there is a wide range of parameters in which DGC dominates both defection strategies in the second model, thereby resulting in high levels of cooperation. Namely, we have shown that, as long as the guilt cost γ satisfies the following condition

$$\frac{T+P-R-S}{2} < \gamma < (\Omega - 1)(R - P), \quad (3)$$

then DGC strategy can dominate all the defective strategies. This condition indicates that, on the one hand, the guilt cost should not be too small in order to ensure guilt has a sufficiently strong effect on emotional players, encouraging guilt alleviation and behavioural change. On the other hand, this cost should not be too large, allowing DGC to compete against unemotional D players who never pay the guilt cost after defecting.

To support the analytical results, we have also provided numerical simulation results, see Figure 1 [5]. Furthermore, those results have

[5] This figure was reproduced from Ref. [16].

been generalised to consider non-extreme or radical guilt modelling (i.e. when $0 < G < \infty$), showing that the obtained results are robust beyond the context of radical guilt strategies (for details see [16]).

Guilt, depending on an agent's strategy, may result in self-punishment, with effect on fitness, and on a change in behaviour. In the first model of guilt, a guilt prone agent is insensitive to whether the co-player also feels guilt on defection. This model does not afford cooperation enhancement because guilt prone agents are then free-ridden by non-guilt prone ones. In our second model, guilt is not triggered in an agent sensitive to the defecting co-player not experiencing guilt too, for instance through telltale signs of eye contact avoidance or frowning (see [19] page 60). It is this latter model that shows the improvement on cooperation brought about by the existence of guilt in the population, and how it becomes pervasive through the usual EGT phenomenon of social imitation. Another successful variation of this model allows to stipulate guilt accumulation coupled with a triggering threshold.

3 CONCLUSIONS AND FUTURE WORK

For sure, we conclude, evolutionary biology and anthropology, like the cognitive sciences too [2, 5, 7, 11, 23], have much to offer in view of rethinking machine ethics, namely for the guilt emotion, evolutionary game theory simulations of computational morality, and functionalism to the rescue [18].

On the basis of psychological and evolutionary understandings of guilt, and inspired by these, this paper proffers and studies two analytical models of guilt, within a system of multi-agents adopting a combination of diverse guilty and non-guilty strategies. To do so, it employs the methods and techniques of EGT, in order to identify the conditions under which there does emerge an enhanced cooperation, improving on the case where there is absence of guilt.

Players evaluate others by their actions of cooperation or defection, whether in the IPD or other models of cooperation. Notwithstanding, they care not simply whether game partners cooperate but pay attention to their decision-making process too. More trust is ascribed to cooperators who have not even considered defecting at all. To quote Kant, "In law a man is guilty when he violates the rights of others. In ethics he is guilty if he only thinks of doing so." [13]. Hence, detecting another's proclivity to cheat, albeit checked by guilt, allots intention recognition an important role to play even when the intention is not carried out [9, 10].

Our results provide important insights for the design of self-organised and distributed MAS: if agents are equipped with the capacity for guilt feeling even if it might appear to lead to disadvantage, that drives the system to an overall more cooperative outcome wherein agents become willing to take reparative actions after wrongdoings.

In future research, the model shall be complicated via our existing EGT models comprising apology, revenge, and forgiveness, by piggybacking guilt onto them [8, 15, 18], namely associating experiencing guilt with joint commitment defection ([24], pp. 108-111).

Last but not least: Currently we only consider one type of emotional strategy playing against unemotional strategy. It is possible that strategies with multiple guilt threshold are co-present in the population. We envisage that different types might dominate in different game configurations, which we will analyse in future work.

ACKNOWLEDGEMENTS

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013. LAMV and TL from Fonds voor Wetenschappelijk Onderzoek - FWO grant nr. G.0391.13N. TL also from Fondation de la Recherche Scientifique - FNRS grant FRFC nr. 2.4614.12. TAH from Teesside URF funding (11200174).

REFERENCES

[1] Bert Brown, 'Face saving and face restoration in negotiation', in *Negotiations: Social-Psychological Perspectives*, ed., D. Druckman, 275–300, SAGE Publications, (1977).

[2] P. Churchland, *Braintrust: What Neuroscience Tells Us about Morality*, Princeton University Press, Princeton, NJ, 2011.

[3] Robert H. Frank, *Passions Within Reason: The Strategic Role of the Emotions*, Norton and Company, 1988.

[4] Benoit Gaudou, Emiliano Lorini, and Eunate Mayor, 'Moral guilt: An agent-based model analysis', in *Advances in Social Simulation*, volume 229 of *Advances in Intelligent Systems and Computing*, 95–106, Springer, (2014).

[5] M. S. Gazzaniga, *The Ethical Brain: The Science of Our Moral Dilemmas*, Harper Perennial, New York, 2006.

[6] Erving Goffman, *Interaction Ritual: : essays in face-to-face behavior*, Random House, 1967.

[7] J. Greene, *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, The Penguin Press HC, New York, NY, 2013.

[8] T. A. Han, L. M. Pereira, F. C. Santos, and T. Lenaerts, 'Why Is It So Hard to Say Sorry: The Evolution of Apology with Commitments in the Iterated Prisoner's Dilemma', in *Proceedings of the 23nd international joint conference on Artificial intelligence (IJCAI'2013)*. AAAI Press, (2013).

[9] T. A. Han, F. C. Santos, T. Lenaerts, and L. M. Pereira, 'Synergy between intention recognition and commitments in cooperation dilemmas', *Scientific reports*, **5**(9312), (2015).

[10] The Anh Han, *Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models*, volume 9, Springer SAPERE series, 2013.

[11] M. D. Hauser, *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*, Little Brown, London, UK, 2007.

[12] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.

[13] Moshe Hoffman, Erez Yoeli, and Carlos David Navarrete, 'Game theory and morality', in *The Evolution of Morality*, 289–316, Springer, (2016).

[14] Stacy Marsella and Jonathan Gratch, 'Computationally modeling human emotion', *Communications of the ACM*, **57**(12), 56–67, (2014).

[15] Luis A Martinez-Vaquero, The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Apology and forgiveness evolve to resolve failures in cooperative agreements', *Scientific reports*, **5**(10639), (2015).

[16] L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, and T. A. Han, 'Social manifestation of guilt leads to stable cooperation in multi-agent systems', in *16th Intl. Conf. on Autonomous Agents and Multiagent Systems*, p. 9 pages. International Foundation for Autonomous Agents and Multiagent Systems, (May 2017 (Accepted)).

[17] L. M. Pereira and A. Saptawijaya, *Programming Machine Ethics*, volume 26 of *SAPERE series*, Springer, 2016.

[18] Luís Moniz Pereira, 'Software sans emotions but with ethical discernment', in *Morality and Emotion: (Un)conscious Journey into Being*, ed., Sara Graça Dias Da Silva, 83–98, Routledge, (2016).

[19] Jesse J. Prinz, *The Emotional Construction of Morals*, Oxford University Press, 2007.

[20] Sarita Rosenstock and Cailin O'Connor. When it's good to feel bad: Evolutionary models of guilt and apology, 2016. working paper.

[21] Karl Sigmund, *The Calculus of Selfishness*, Princeton University Press, 2010.

[22] TheEconomist. March of the machines - a special report on artificial intelligence, June 25, 2016.

[23] M. Tomasello, *A Natural History of Human Thinking*, Harvard University Press, Cambridge, MA, 2014.

[24] Michael Tomasello, *A Natural History of Human Morality*, Harvard University Press, 2016.