Cooperating AI

Making artificial intelligence more human

Julian L. Heiss

Ist künstliche Intelligenz gefährlich?
Seminar Report
Supervision: Dr. Ulrich Köthe
Heidelberg University

September 2017

1 Introduction

Cooperation amongst humans is without doubts a thriving factor in the evolution and progress of this species. When talking about humans, one has to talk about the concept of society, which cannot the thought of without coperation.
It is also undeniable that man-made machines are becoming more and more important for the human life and the current standard of living is highly dependent on machines and algorithms that are guiding these machines through their tasks.
As the algorithms evolve from a simple rulebook to sophisticated programs that exhibit characteristics that can be described by the notion intelligence, new possibilities of applications arise. Not only do these machines need less supvervision, but they can also be interacted with in different manners than just giving orders.
Since these machines are designed by different people than which are using them, it is common that the machines set preferences are not shared by the user. Therefore it is obvious to want the machine to be able to cooperate with different human users in a wide variety of situations.
While there have been huge steps in the ability of machines to perform cognitive tasks, the development of cooperative behaviour has been less investigated, since it is less dependent on computational power and the objective of creating cooperative behaviour is less accurately defined than most cognitive tasks in the field of artificial intelligence.
As the range of problems that machines have trying to learn cooperation is wide, so is the range of potential solutions. The set of possible solutions also grows bigger, as the artificial intelligence (AI) evolves. While there is no single strategy that proves to be unversally applicable, some ideas about cooperation between humans and machines are discussed in the following.

2 Multi-Agent Reinforcement Learning in Sequential Social Dilemmas

In [1] Leibo et al. looked at Machine-Machine cooperation through analyzing the behaviour of learning agents in (Markov) games that model social dilemmas.

Game-theoretical set-up Leibo et al. argue that the status quo of using repeated general-sum matrix games (e.g. like Prisoner's Dilemma, Chicken, and Stag Hunt) as a framework for understanding social dilemmas has some drawbacks with regards to modeling real world social dilemmas [1], such as they are neglecting the temporal dimension and that they do not look as cooperation as a graded quantity. Therefore they propose instead the framework of sequential social dilemmas (SSDs), that extend the existing framework, while still exhibiting the mixed motivation structure of either choosing cooperational or defective behaviour.
As examples of SSDs they considered the Gathering and Wolfpack games, which both derive from the Prisoner's Dilemma in the old matrix game social dilemma framework, but

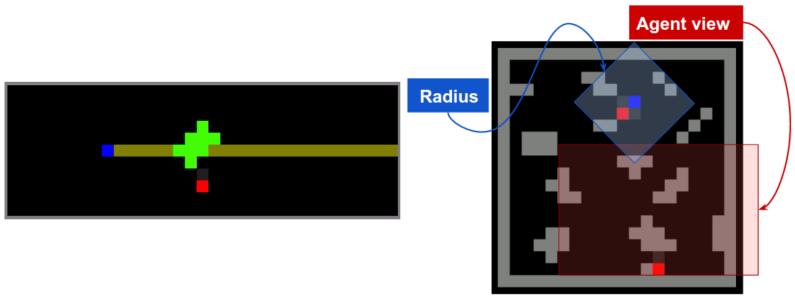

Figure 1: The acting agents in the Gathering game (left) are blue and red, while the apples are depicted in green. For the Wolfpack game (right) the agents are both red picels, while the prey is colored blue. Taken from [1].

exhibit different properties in the new framework. So different that small experimental changes even yield opposite results for both games.

Objectives In the scope of this seminar and with the goal of understanding cooperation of artificial intelligence, rather than understanding the actual framework and game theoretical traction of these games, it is more important to gain some intuition from this experiment.
The question raised by the authors is: "What social effects emerge when each agent uses a particular learning rule?". So this work aims at characterizing the dynamics resulting from the learning rules and not at designing new rules.
The learning rules and games investigated here are chosen since the authors claim that this class of reinforcement algorithms is seen as a candidate theory of animal habit-learning [1].

Set-up of the games The games, shortly explained, are set up as follows: Each agent has only a partial of the surrounding environment. An agent must learn a policy while coexisting with one another. Depending on its impact on the other agent, a policy is either cooperational or a defection policy.
In the Gathering game the agents are rewarded for collecting "apples" (green pixels), which respawn at a specified time after having been collected. It is also possible to "shoot" a beam onto the other agent, an doing this twice removes the agent hit - or "tagged" - by the beams for a different specified respawn time. No reward is given for tagging the other player, the only motivation is competition over the apples.
In the Wolfpack game, both players - the wolves - chase a prey. When a wolf "touches" the prey, all wolves within a "capture radius" receive a reward which is proportional to the amount of wolves within the radius. The idea is, that two wolves can better protect the carcass from scavengers after catching the prey.

One might refer to figure 1 or watch the gameplay videos for better understanding[1] [2].

The agents are trained using deep reinforcement learning methods, where the cumulative long-term award is to be maximized by the agents by repeatedly playing instances of the games. The individual agents are trained with a deep Q-network. They learn independently of one another, only being considered by the other player as part of its environment.

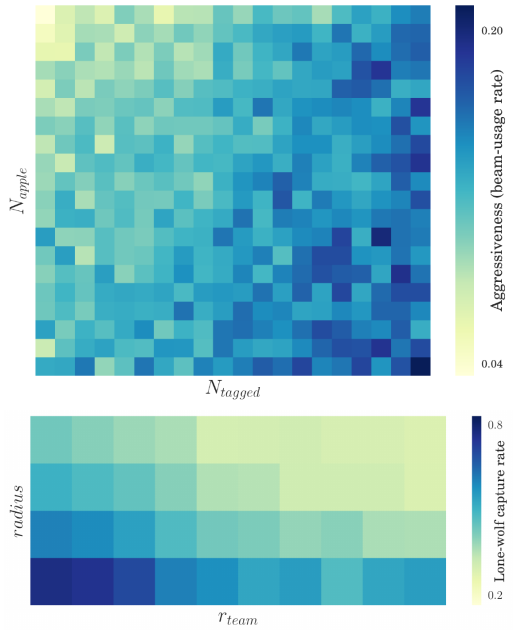

Figure 2: Influence of environmental parameters on the evel of cooperation in the Gathering game (top) and the Wolfpack game (bottom). For both graphs a brighter colour represents a higher level of cooperation. Taken from [1].

Evironmental influence Leibo et al. show that, in the Wolfpack game, increasing the bonus of catching a prey together (r_{team}) as well as increasing the capture radius both increase the level of cooperation as seen in figure 2. This shows the obvious impact of the environmental parameters in the games on the level of cooperation, which is measured by the inverse of the rate of successful lone-wolf captures.

In the Gathering game it is intuitive to measure the level of cooperation by the usage rate of tagging action. A high usage rate indicates an agressive defecting policy. As in the Wolfpack game, again the environmental parameters - the abundance of apples (N_{apples}) and the respawn time after being tagged (N_{tagged}) - were affecting the level of

[1] `https://goo.gl/2xczLc`
[2] `https://goo.gl/AgXtTn`

cooperation, as can be seen in figure 2. This experiment predicts the emergence of conflict in an environment with scarce resources.

Agent parameters For the Wolfpack game, learning the "lone-wolf" policy is easier than learning a cooperative "pack-hunting" policy. This is intuitively understandable because the the lone wolf does not need to include neither the other agent's behaviour nor the existence of a capture radius into consideration for his actions. Greater network size then leads to more cooperation since it allows the agent to take these complications into consideration.
In the Gathering game the situation is reversed. Cooperative policies are easier to learn since they need only be concerned with apples and may not depend on the rival player's actions. Peaceful coexisting is an easy form of cooperation as long as the environmental parameters allow this behaviour. For Gathering, an increase in network size leads therefore to an increase in the agent's tendency to defect, since due to the aiming involved in tagging, the defecting policy is a more complex task.
The authors state that these kind of qualitatively differences support the need of the SSD framework for the modeling of real social dilemmas.

Conlusion From these two examples it can be seen that there is no easy and obvious relation between cognitive capacity and cooperative behaviour. I.e. increasing capacity does not automatically make an the algorithm more cooperative.
Also, since cooperation is studied, one has to be aware that shooting a beam might still be favourable in the Gathering game, e.g. so that not both agents go for the same apple. There exists the problem that a chosen set of rules might give unwanted incentives. Incentives that the creator of the reward functions was not aware of and some of which can cause problems in situations where humans are dependent on the behaviour of the robot agents.

There is still an obvious need to improve cooperation in these games. A possible way to achieve this would be learning the reward function for the game before implementing it by mimicking humans or using instructive teaching (see section 4). Another possibility is giving the agents the ability to communicate, an idea which has been investigated by Crandall et al. in [2] and will be discussed in the next section.

3 Cooperating with machines

In the previous section we just considered machine-machine cooperation in special areas. But for machines to be cooperating with human agents, the used algorithms have to capture the concept of cooperation in broader sense.
The goal of the research of Crandall et al. in [2] is to create an AI algorithm that is cooperating with machines or humans at the level of human cooperation. To be able to

understand and quantify the experiment's outcome this was done in arbitrary two-player repeated interactions.
There are several conditions that an algorithm has to fulfil to be considered successful:

- Generality: It has to function in a wide variety of scenarios.
- Flexibility: It has to function with both humans and machines without knowing the partners behaviour a priori.
- Learning Speed: It has to be able to learn cooperative behaviour in only a few rounds, specifically to accomodate for the human timescale.

A variety of standard machine learning algorithms, belief-based algorithms and expert algorithms could not produce the results with regards to cooperation that the research group was hoping for. The best performing algorithm was S++ [3].
A hypothesis for the algorithms lack of ability to form "effective long-term relationships" with people or other machines was that the agents could not coordinate their actions properly since they had no way to communicate their actions and intentions.
The idea of Crandall et al. was therefore to extend one of the algorithms by the possibility of communication between the acting agents.
This ability of communication comes natural to humans, however, this can not be said about algorithms. The strength of some algorithms, like neural networks, is even based on their inherently different structure to the human thought and decision process. For these kind of algorithms it would not be trivial to express their intent in a representation that humans understand.
The aforementioned, best performing algorithm S++ has luckily a understandable high-level representation of its strategy, as it is an expert algorithm, that selects one of a finite set of strategies. The expert strategy that the algorithm chooses at each round can be communicated as the intention of its next action.
The idea of Crandall et al. in [2] was to extend the algorithm S++ by the element of communication to create a new algorithm that is better suited for fulfilling the aforementioned requirements of an cooperative algorithm.
More specifically, they allowed for communication via cheap talk, which refers to "non-binding, unmediated, and costless communication" [2]. This helps to create shared representations for the players, as it allows for mutual feedback and planning during the game.

3.1 Experiment

Set-up Crandall et al. conducted a study in which participants played three representative repeated games, which were drawn from distinct pay-off families (Chicken Game, Alternator Game, and Prisoner's Dilemma).
The participants played via a computer interface that was obscuring the identity of their opponent.
To investigate the effect of cheap talk, some players could send messages at the beginning

of each round via the computer interface. These messages were confined to a set of 19 different speech acts (listed in figure 3) that the algorithm could learn to use.

Speech ID	Text	Speech ID	Text
0	Do as I say, or I'll punish you.	10	We can both do better than this.
1	I accept your last proposal.	11	Curse you.
2	I don't accept your proposal.	12	You betrayed me.
3	That's not fair.	13	You will pay for this!
4	I don't trust you.	14	In your face!
5	Excellent!	15	Let's always play <action pair>.
6	Sweet. We are getting rich.	16	This round, let's play <action pair>.
7	Give me another chance.	17	Don't play <action>.
8	Okay. I forgive you.	18	Let's alternate between <action pair>
9	I'm changing my strategy.		and <action pair>.

Figure 3: Permitted speech acts for communicating with the partner in the experiment. The machine-machine pairs also used these when cheap talk was permitted. Taken from [2].

Algorithm A graphical overview explaining the decision-making process of S# as well as its process of choosing a speech act can be found in [2, Fig 1].

Measuring cooperation In the scope of the discussed paper, the level of mutual cooperation is defined as the Nash bargaining solution [4] of the games or the solution that maximises the product of the advantages of the players. The mutual cooperation is therefore defined differently for the different games that are investigated in this work [2, Supplementary Information].

3.2 Results and Properties

Effect of Cheap Talk The obvious and most important result from the experiment is the overall improvement of cooperation when communication is allowed. This can be seen as a clear trend in figure 4 as all possible pairings score higher on the cooperation metric with the match-ups that involve humans almost doubling their cooperation level.

Loyalty In the following we see indication that not only machines struggle with cooperation, but also humans themselves. They just struggle with a different part of cooperative behaviour: To forge mutually cooperative relationships, players must do two things: Establish cooperative behaviour and maintain it.
One can see in figure 5 how the cheap talk is helping the player to establish cooperation. Once a successful strategy has been found, the machine agents have no tendency to deviate from that strategy.
"Loyalty" (to the strategy) can therefore be named as one reason for M-M pairs outperforming humans in this experiment.

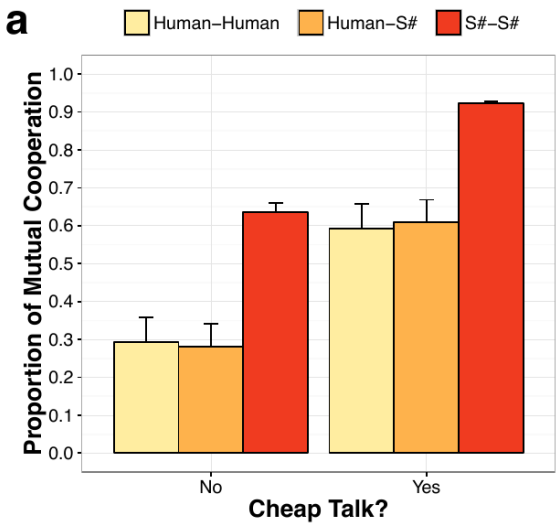

Figure 4: Effect of cheap talk on cooperation behaviour. Taken from [2].

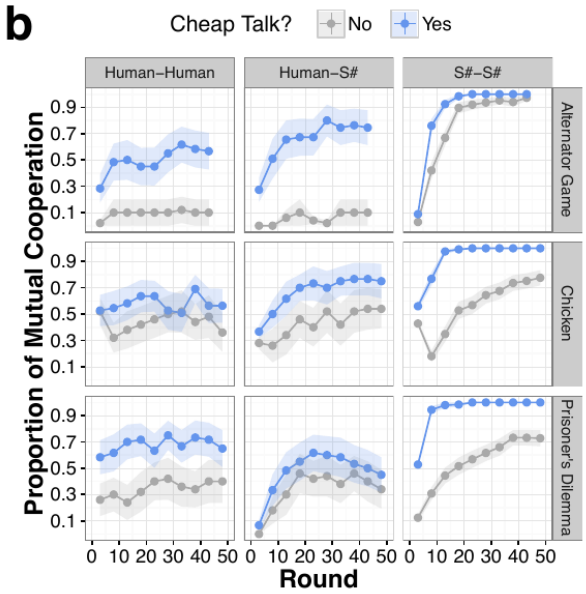

Figure 5: The average proportion of mutual cooperation over time in each game in each pairing and condition. Picture and caption taken from [2].

Honesty Another reason for machine pairs outperforming humans is "honesty". Since verbal commitments by S# are derived from its intended behaviour, it does what it says, unlike a significant portion of the human participants, as can be seen in figure 6. Exploring the data even further, figure 7 looks at the potential gains of cooperation the H-H and H-M could have had if they had been loyal and honest throughout the experiment.
It is also important to notice that they did not only lose potential cooperation but also

gained less reward in all but two cases. Which means that this behaviour did not even make sense from an purely egoistical viewpoint.

out of how many?

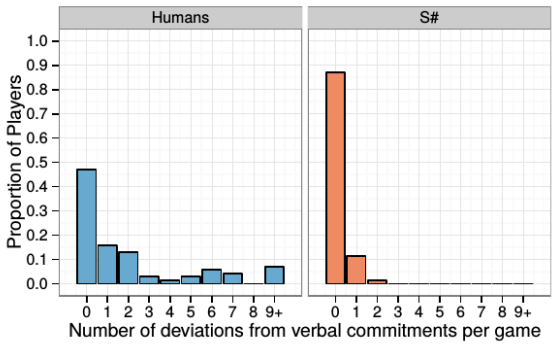

Figure 6: A histogram showing how often players deviated from their verbal commitments during the course of a game. Deviations from the proposed plan were not counted (1) if the player instead followed a proposal made by its partner or (2) after the player's partner deviated from the proposed plan. Picture and caption taken from [2, Supplementary Information].

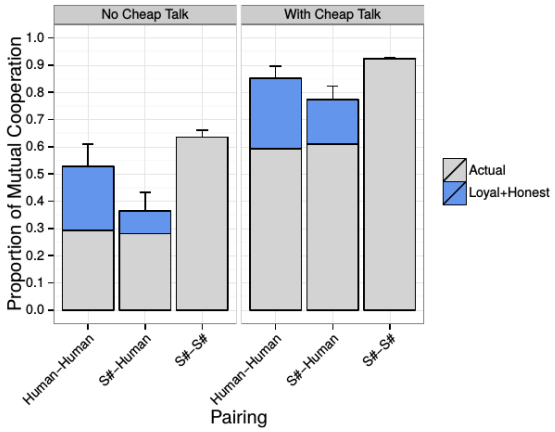

Figure 7: The estimated proportion of rounds that would have resulted in mutual cooperation had all human players followed S#'s learned behavioral and signaling strategies of not deviating from cooperative behavior when mutual cooperation was established (i.e., loyalty) and following through with verbal commitments (i.e., honesty). Had all human participants been loyal and honest, these results indicate that there would have been little difference between Human-Human and S#-S# pairings. Picture and caption taken from [2].

Speech Profile Differences between human and machine players can also be seen in the usage of the available speech acts which is displayed in figure 8. The machine players

used a significantly greater amount of threats and hate speech, while the human players were more often praising their partner. But graph does not necessarily imply that the AI is more "evil" than humans. Maybe threats are just a more "effective" way to ensure cooperation.

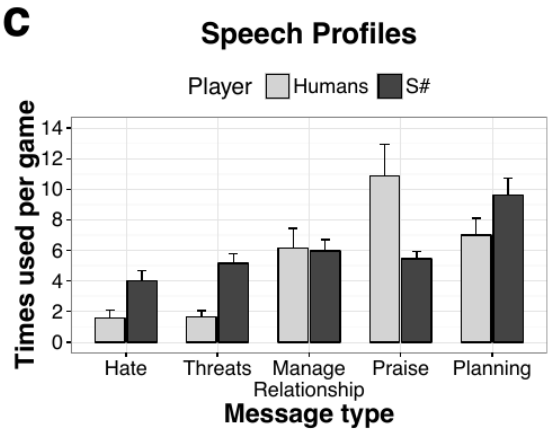

Figure 8: The average number of times that Humans and S# used messages of each type over the course of an interaction when paired with people across all games. The permitted speech acts are grouped in 5 categories. Picture and caption taken from [2].

3.2.1 Conclusion

While the authors of this work claim that "The machine-learning algorithm learned to be loyal." (J. Crandall [9]), my personal impression is, that the exhibited loyalty derives more from an lack of ability to betray than a learned capability.
Yet they succeeded in creating a more cooperative algorithm by mimicking humans and their favourite way of organising cooperation: Communication.

4 Cooperative Inverse Reinforcement Learning

As we noticed in the first section with the Wolfpack and Gathering games, differences in the reward functions of the agents can be crucial to the success of the cooperation. Also, obviously, it is very difficult to program this reward function by hand without leaving room for any misinterpretations by the robot agent.
A common example for the possible misinterpretation is a robot, that is supposed so vacuum a room. If rewarded for cleaning up dirt, the optimal policy that the robot can follow is repeatedly dumping and cleaning up the same dirt [5].

≪If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively [...] we had better be quite sure that the purpose put into the

machine is the purpose which we really desire.» (Norbert Wiener, 1960) [5].
This quote refers to the **value alignment problem**, which roughly states that highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation [6].

In general, **inverse reinforcement learning** (IRL) is used to infer an agent's reward function by observing his behaviour. Normally the agent is considered to be acting optimal or close to optimal, since otherwise it would not make much sense to mimic his behaviour. IRL seems like a good solution to the value alignment problem, as making the machines mimic the human behaviour is supposed to ensure that their values are aligned. However, there are some problems to that, as Crandall et al. point out:

- With IRL, the robot learns the human reward functions, but applies it to himself. Instead the robot should have the objective of optimizing the reward **for** the human.
- IRL assumes, as stated above, that the observed behaviour is optimal and this is not in general the most efficient way of teaching. Efficient teaching may involve more interaction than letting somebody watch what you are doing. And this extra bit of interaction is not included in the IRL framework.

To tackle these problems, Hadfield-Menell et al. proposed to extend the IRL framework to create a cooperative and interactive reward maximization process called **cooperative inverse reinforcement learning** (CIRL).
To be more accurate, a CIRL problem is a cooperative partial information game involving two agents, human (H) and robot (R). Both are rewarded according to the human's reward function, but the robot does not initially know what this is.

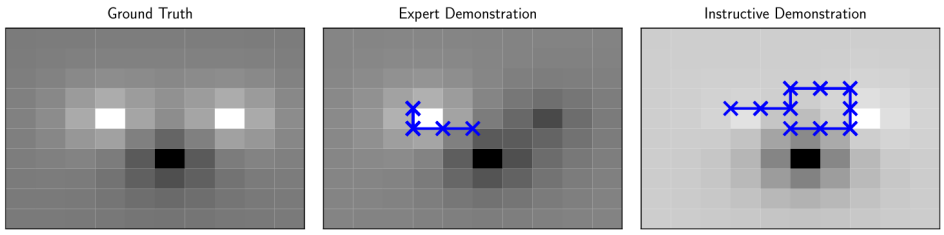

Figure 9: The difference between demonstration-by-expert and instructive demonstration in a mobile robot navigation problem. The backgrounds show the reward function, lighter grid cells indicate areas of higher reward. The left panel shows the ground truth that should be learned. The middle and right panel show the learned functions after the expert and the instructive demonstration respectively. Superimposed in blue is the path used by the instructor. Picture taken from [5].

The difference between CIRL and IRL can be intuitively explained by considering the example from figure 9. The demonstration manages to highlight both panels of high reward, while the expert policy successfully teaches the location of maximum reward, but

misses to teach the second maximum. Therefore the robot receives a better estimate by the instructive demonstration which does not fulfil the assumption of optimality.

The contribution of Hadfield-Menell et al. in [5] consists in presenting a new game-theoretic model for cooperative learning, where the robot knows that it is learning to maximise the human's reward. They show that in this model the task of finding an optimal policy pair can be reduced to solving a POMDP (partially observable Markov decision process).
With respect to the value alignment problem Hadfield-Menell et al. state that the following: "Returning to Wiener's warning, we believe that the best solution is not to put a specific purpose into the machine at all, but instead to design machines that provably converge to the right purpose as they go along." [5]

Even with this improved version of inverse reinforcement learning one can not claim the value alignment problem to be solved. This is due to the reason that the problem itself might not be well stated. It is put as the problem of aligning the machines values to our values. But it is not a priori clear what our values are - what "we" "want". It is a big challenge for AI research due to several reasons:

- It is not easy to encode human values in a programming language [6].
- Humanity does not (yet) agree on common values.
- The values that are agreed on are not set in stone and might change with time.
- Since human values are variable it is not assured that they are the best values there can be. And also by which measure this has to be decided.

So the problem of value alignment can not be solved without solving the current moral, political sociological conflicts and ultimately solving the question what 'human values' are [8].

5 Friendly AGI via Human Emotion: the Vital Link

The following section is based on Dietsch's work in [7] and discusses ethical decision-making by Artificial General Intelligences (AGI) that are implemented as meta-beings consisting of individual components and shared data. Data which might also include human data.

One has to be aware that rules and sets of values are not sufficient to guide artificial intelligence as some situations involve trade-offs. Scenarios which are often brought up, when talking about artificial intelligence, autonomous systems and robot ethics, are dilemmas of the kind of the railway siding problem, where one has to choose between two subjective evils with negative outcome. Human intercession in this kind of scenario is clearly

not feasible, since humans would not be able to respond at the AI's timescale nor is it obvious how an AGI even becomes aware that a situation calls for an ethical action or decision [7]. Recognition of a problem is therefore the first step of arriving at a decision.

The decision making of an AGI is obviously different to the human one, but it is prone to the same constraints:

- Machines (also quantum computers) will always have some limit on computational capacity. This means there will always be a information overflow. Distillation and filtering of information is therefore an inevitable part of the decision making process of an AGI, just as it is for humans.
- These filtering processes are based on memory, pattern recognition, prediction, evaluation methods, all of which are complex processes. This forces a prioritization of the tasks and makes AGIs as well as humans, when confronted with a problem, follow prescribed methods in the decision making process to avoid extra computational costs.

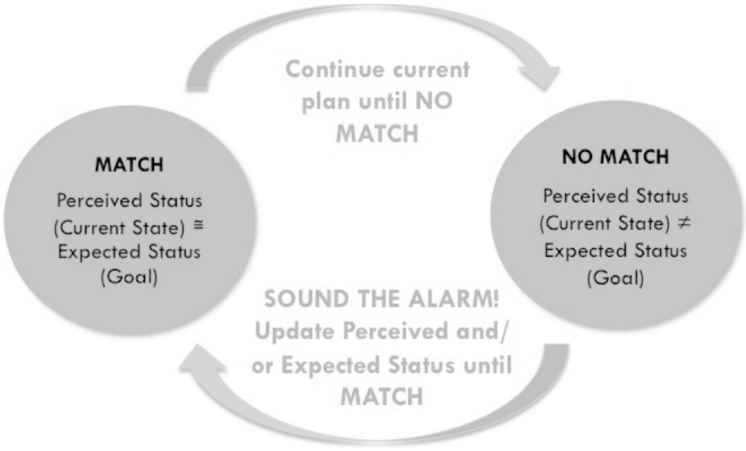

Figure 10: The human ACC lets habit proceed until it detects a mismatch between Perceived State and Expected State. Picture and caption taken from [7].

Difference Engine For humans, the Anterior Cingulate Cortex serves as a "difference engine" that compares expected states with the current perceived state.
The prioritization of situations is done by valuating the difference between the expected and perceived state. Dietsch exlpains how humans use emotions to drive actions to achieve homeostasis of their needs (like physiological/safety needs, social/esteem needs and self-actualization needs).
AGIs, will have similar needs. However, their physical needs are simple and probably

easily achieved. Safety concerns could be potential data breaches. And cooperative AGIs, the ones we are specifically interested in, are also expected to have some social needs.

Proposal It is now important to notice that social and safety needs require a distinctions between self and others. Dietsch now argues that it is of upmost importance, that humans are innate members of the "ingroup" of the AGI. Even going so far of suggesting, that "we" and "me" should be inseperable for the AGI will always include the human well-being in their decision-making process. Its expected and perceived states have to icnlude the human data.

She also states arguments against linking human emotion and AGI in "Meta-Beings", such as loss of privacy, freedom an individuality, as well as the question of who dominates and whose needs dominate.
The latter at least is countered by the notion, that our needs would be part of the self of the AGI and therefore are always considered.

Conclusion This is a more speculatvive work, with no ultimate solution to ethical decision-making by AGIs, but it is giving the starting point for one.
The suggestion is linking the humans emotions and perception to the AGI and its wellbeing as a reliable way to assure the AGI's attention to the well-being of humanity - whatever the actual implementation might be [7].
Obviously this is an interesting concept, although the implementation is so unclear that it is highly speculative and some of the arguments against the proposed linking will stick around for much longer. It is an concept that should be remembered and considered when it is clearer how machines and humans will interact in the future and when it is foreseeable if AIs can evolve to AGIs.

6 Conclusion

Throughout the discussion of the selected scientific work, we have seen the importance of communication for cooperational behaviour.
Another way of increasing the ability of machines to exhibit cooperational behaviour is to use existing mechanisms and algorithms, like inverse reinforcement learning, to introduce concepts as human values into an artificial intelligence.
Altough these approaches yielded some positive results, this might not be sufficient to solve problems as the value alignment problem. To overcome this, a linking between human emotion and artificial perception has been proposed, but this suggestion remains highly speculative for the near future.

However, it is the correct appraoch to not inly develop the cognitive capacities of AIs further, but one also has to be aware of the part that the human species itself has in the value alignment problem. AI research will have to collaborate and cooperate with social sciences to tackle these kind of problems, where the general objective is not well-defined.

References

[1] Leibo et al. (2017). Multi-agent Reinforcement Learning in Sequential Social Dilemmas. Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AA-MAS 2017).

[2] Crandall et al. (2017). Cooperating with Machines. Computing Research Repository (CoRR), abs/1703.0. http://arxiv.org/abs/1703.06207

[3] J. W. Crandall. Towards minimizing disappointment in repeated games. Journal of Artificial Intelligence Research, 49:111-142, 2014.

[4] J. F. Nash. The bargaining problem. Econometrica, 28:155-162, 1950.

[5] Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S. (2016). Cooperative Inverse Reinforcement Learning, (Nips). Retrieved from http://arxiv.org/abs/1606.03137

[6] https://futureoflife.org/ai-principles/
Ariel Conn.
https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/

[7] Dietsch, J. (2014). "Friendly" AGI via Human Emotion: the Vital Link. AAAI 2014 Fall Workshop.

[8] http://duckofminerva.com/2017/01/the-value-alignment-problems-problem.html

[9] https://www.sciencemag.org/news/2017/03/computers-learn-cooperate-better-humans