Hybridized Assistance Games and Value Alignment

Frank Wildenburg (6263585)
Supervisor: dr. Natasha Alechina
Second evaluator: dr. Janneke van Lith

Abstract

A fairly recent proposal in the study of value alignment is the assistance game, in which initially unsure agents learn to maximize human preferences by observing human behaviour. Here, we propose that assistance game-based agents might benefit from being "hybridized" with other AI techniques. To describe these hybridized systems, we first consider the advantages and disadvantages of assistance games, before considering in what ways a hybridized agent may work and how an assistance game-based agent with sufficient computational resources might be motivated to create a hybridized system by using other AI technique(s). To illustrate the beneficial effects of a hybridized system, we consider ways the effects of these systems might fulfill the requirements of trustworthy AI described by the European Union's High Level Expert Group on Artificial Intelligence.

Keywords: Value alignment, assistance games, trustworthy AI

A 15 ECTS thesis submitted in fulfillment of the requirements of the Bachelor of Science in Artificial Intelligence at Utrecht University, submitted on the 27th of November, 2020

With my sincere thanks to all the people who assisted me throughout this period, whose number is as large as the amount of help they gave

Contents

Introduction

There is an increasing presence of artificial agents that interact with humans and the world around them to achieve various goals, typically in the interest of individual or societal well-being. This presence has led to significant positive effects on transportation, health, climate science and many other areas (Russell & Norvig, 2020, Chapter 1).

However, there is also an increased awareness of the fact that the effects of artificial agents are not necessarily positive. In recent years, one can look at recommender systems maximizing click-through by changing the preferences of visitors - rather than by recommending relevant content - and in doing so leading to polarisation (Alfano, Fard, Carter, Clutton, & Klein, 2020). On a larger timescale, scenarios have been sketched in which artificial agents chasing seemingly harmless goals cause existential threats, such as Bostrom's (2003) paperclip maximizer.

The question of how to achieve successful collaboration between AI systems and humans has been formalized in AI safety literature as the *value alignment problem* (Amodei et al., 2016). Bostrom (2014) broadly describes three possible approaches to value alignment: to start out with an agent already loaded with commendable values (such as might happen when whole-brain emulation is achieved), to directly specify the values we want the agent to have into its programming, or to create some mechanism by which the agent can obtain values based on its environment.

Due to the recent developments in the field of artificial intelligence, much research has been done into the second and third of these approaches. Specifying the values can be done in many different ways, ranging from norm-based approaches in which many different values are specified to machine learning methods in which the single function which the agent should optimize represents a single value (Russell & Norvig, 2020).

However, in his recent book, Russell argues that any approach which uses values specified by humans is inherently flawed, due to the fact that humans do not know all details of the values they have (Russell, 2019). Specifying a seemingly commendable value, therefore, might lead to detrimental effects in some unconsidered situation. Instead, a better approach uses the following three principles:

1. The only goal of artificial agents is maximizing human preferences
2. The agent should be initially unsure about these preferences
3. The best source of information about these preferences is human behaviour.

Using these principles, Russell proposes to use the concept of *assistance games*. In these assistance games, agents attempt to fulfill human preferences that they themselves are possibly unsure of. When unsure about the positive effects of a plan, the agent can acquire extra information by allowing humans to make a decision to either allow the agent to execute its plan or to turn off the agent. Eventually, this information will allow agents to reach points at which they are confident enough to start their plans without consulting humans, while having values that match those of the humans they learned from.

According to Russell, agents based on these assistance games have several benefits: they enable safe interruptibility, ensuring the agent does not attempt to avoid being turned off when a human wants to do so; they help avoid wireheading, in which the agent attempts to adjust the reward signal which they receive; and agents based on assistance games might be likely to avoid dangers in cases of recursive self-improvement.

While the arguments for these advantages of assistance game-based agents are strong, we also see some possible dangers with agents based on these principles. Specifically, risks exist with more complex questions with regards to preferences, such as what to do in cases of preferences that are contradictory either in a person or between multiple persons; if agents maximizing preferences should always obey the orders they are given; and if agents will always behave the same way in cases without a supervisor as in cases similar expect with a supervisor.

To counter some of these disadvantages of assistance game-based agents, we propose and describe agents based on a hybridized form, in which traditional system designs in which values are specified are combined with assistance games. To show the possible benefits of such agents, we describe possible scenario's in which advantages might be gained. To do so, we use the Ethics Guidelines for Trustworthy AI published by the European Union's High Level Expert Group on Artificial Intelligence.

The rest of this paper is laid out in the following way. First, we will describe assistance games and their advantages and disadvantages, showing how these (dis)advantages might express themselves with the help of a case study. Using these facts, we will describe a way in which a hybridized form of artificial intelligence, using both assistance games and traditional systems, can be implemented. We will then show how such a hybridized agent could avoid the detrimental effects shown in the case study, and describe how such hybridized systems can be used to achieve the requirements of trustworthy artificial intelligence described by the Ethics Guidelines for Trustworthy AI. Finally, we conclude and describe possible directions for further research.

Multi-disciplinarity and social context

Although this research is situated first and foremost within the field of artificial intelligence, this does not mean that it can be seen as monodisciplinary. Artificial intelligence has, after all, always been situated between the humanities, the social sciences, and the exact sciences. As such, any research into artificial intelligence, and especially into more fundamental aspects of artificial intelligence such as optimal agent design, can be approached as inherently multi-disciplinary. In a sim-

ilar manner, the need to integrate theoretical artificial intelligence with concrete policies made by institutions such as the European Union or UNESCO requires a socially conscious approach to research. It is for these two reasons that this thesis can be seen as meeting the requirements for a Humanities Honours Thesis.

Assistance Games

Assistance games, also known as cooperative inverse reinforcement learning (CIRL) games, are based on the older field of *inverse reinforcement learning* or IRL. An IRL algorithm attempts to determine the reward function of an agent by observing the actions that agent takes (Ng & Russell, 2000). These actions are assumed to be approximately optimal.

Hadfield-Menell, Dragan, Abbeel, and Russell (2016) describe two flaws in assuming that an IRL algorithm provides a simple solution to the value alignment problem. Firstly, we need to ensure that the reward function the agent adopts is not simply the human reward function - which might lead to the weird situation in which a robot taught by someone who enjoys coffee wants coffee for itself - but rather an objective of optimizing the reward for the human based on the reward function it learns. Secondly, the assumption that the actions observed by the robot are approximately optimal excludes the possibility of useful teaching behaviours, which would be desirable in situations in which agents are expected to quickly learn new tasks.

To avoid these two problems, Hadfield-Menell et al. define a cooperative inverse reinforcement learning game as a two-player game in which the robot's payoff is the human's actual rewards, the function of which is known by the human but not by the robot. This allows value alignment to be formulated as a cooperative and interactive process. The structure of CIRL games also allows the computing of optimal policies for human and agent to be reduced to a single-agent partially observable Markov decision process.

Research has also been done into situations in which the agent is influenced by multiple humans, in a variation called the multi-principle assistance game (Fickinger, Zhuang, Hadfield-Menell, & Russell, 2020). In these cases, there exists the possibility that it is not possible to have the agent perfectly match the preferences of everyone simultaneously, even if that would be possible for everyone individually. Furthermore, measures should be taken to ensure that agents do not misrepresent their preferences to gain a more desirable outcome. Fickinger et al. look at these problems from the perspective of social choice theory (Sen, 1986), while Russell (2019) mentions that agents should consider the preferences of multiple humans while not simply abandoning the owner of the agent.

Formalisation (Hadfield-Menell et al., 2016)

Formally, a cooperative inverse reinforcement learning game is a game with identical payoffs between a human $\mathbf{H}$ and a agent $\mathbf{R}$ described by a tuple $M = \langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}, T(\cdot|\cdot,\cdot,\cdot), \{\Theta, R(\cdot,\cdot,\cdot;\cdot)\}, P_0(\cdot,\cdot), \gamma \rangle$ where:

$\mathcal{S}$ is a set of world states: $s \in \mathcal{S}$

$\mathcal{A}^{\mathbf{H}}$ and $\mathcal{A}^{\mathbf{R}}$ are sets of actions for $\mathbf{H}$ and $\mathbf{R}$: $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}}$ and $a^{\mathbf{R}} \in \mathcal{A}^{\mathbf{R}}$

$T(\cdot|\cdot,\cdot,\cdot)$ is a conditional distribution on the next world state, given previous state and action for both agents: $T(s'|s, a^{\mathbf{H}}, a^{\mathbf{R}})$

Θ is a set of possible static reward parameters, only observed by $\mathbf{H}$: $\theta \in \Theta$

$R(\cdot,\cdot,\cdot;\cdot)$ is a parameterized reward function that maps world states, joint actions, and reward parameters to real numbers $R : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \Theta \to \mathbb{R}$

$P_0(\cdot,\cdot)$ is a distribution over the initial state, represented as tuples: $P_0(s_0, \theta)$

γ is a discount factor: $\gamma \in [0,1]$

The game starts by sampling the initial state from P_0. $\mathbf{H}$ observes θ, but $\mathbf{R}$ does not. Then, at each timestep t, $\mathbf{H}$ and $\mathbf{R}$ observe the current state s_t and select their actions $a_t^{\mathbf{H}}, a_t^{\mathbf{R}}$. Both actors then receive rewards $r_t = R(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{R}}; \theta)$ and observe each other's action selection. Finally, a state for the next timestep s_{t+1} is sampled from the transition distribution, and the process repeats.

The action selection for $\mathbf{H}$ and $\mathbf{R}$ are determined by a pair of policies $(\pi^{\mathbf{H}}, \pi^{\mathbf{R}})$. The optimal joint policy is the policy that maximizes *value*, where the value of a state is the expected sum of discounted rewards under the initial distribution of reward parameters and world states.

Hadfield-Menell et al. (2016) also prove that, given an arbitrary CIRL game with state space $\mathcal{S}$ and reward space Θ, there exists a POMDP M_C with hidden state space $\mathcal{S}_C$ such that $|\mathcal{S}_C| = |\mathcal{S}| \cdot |\Theta|$ and that, for any policy pair in the CIRL game, there is a policy in M_C that achieves the same sum of discounted rewards. Furthermore, these POMDPs, while still possibly very challenging, are less computationally complexive than the NEXP-complete decentralized POMPDs that are necessary to compute the optimal joint policy for a general cooperative game. Additionally, the structure of these POMDPs enables more efficient algorithms (Russell & Norvig, 2020, Chapter 18.2.5).

Using this system, we can create a mechanism in which the learning done by artificial intelligence is very similar to learning done by humans in training: instead of simply being told or shown how a successful action is performed, the agent is explained in detail how the task is performed, including information on what to do in edge cases or in situations where things went wrong and have to be fixed. which enables it to learn what is truly expected of it. Due to this fact, there are several advantages to this method of specifying values.

Advantages of assistance games

The first of these advantages, and one of the major reasons to advocate the development of agents based on assistance games, is the fact that the values of the agent do not have to be manually specified by some human programmer or researcher. The danger inherent in this manual specification of

seemingly desirable values or properties can already be seen in the tale of king Midas, as well as Asimov's *Robot* series of books, where seemingly beneficial wishes or goals cause detrimental effects.

Nevertheless, several seemingly promising attempts have been proposed in AI research. For example, Schmidhuber (2007) proposes that a desire for discovery and beauty can be encouraged by maximizing the measure "create action sequences that extend the observation history and yield previously unknown / unpredictable but quickly learnable algorithmic regularity or compressibility", while Bostrom (2014) suggests the possibility of specifying processes of deriving a standard by defining the final goal of the agent as something along the lines of "achieve that which we would have wished the AI to achieve if we had thought about the matter long and hard.".

Unfortunately, problems can be identified with these goals as well. For example, one way to meet Schmidhuber's goal is to present the agent with a long series of regular data encrypted in some complex way, and then reveal the secret of this encryption to the agent, allowing for a large compression on its past data. Bostrom's suggestion, meanwhile, might lead to trouble if the agent assumes that thinking sufficiently long about the matter would lead to the questioner passing away from old age, leaving them with no wishes for the AI.

Nor, argues Soares (2018), would we be successful by repeatedly patching the flawed goals. By patching the possibility that allows one forbidden pathway, we ensure that the agent will follow the nearest non-forbidden pathway. Since there are infinitely many of these pathways, it is not possible to patch every flawed goal in this manner.

For this reason, the fact that agents based on assistance games have no need for human-specified values allows them to avoid situations not specifically specified as forbidden but considered extremely undesirable. As long as the behaviour of the humans the agent learns from is considered acceptable, the agent could even be used in situations where no consistent codification of ethics exists.

A second considerable advantage of agents based on assistance games is the safe interruptability of these agents. This, formalized by Orseau and Armstrong (2016), refers to the ability of humans to safely interrupt an agent while making sure the agent does not attempt to learn to either prevent or induce these interruptions.

While intuitively it might seem that there is no reason to implement features that would lead to agents attempting this, Omohundro (2008) argues that self-preservation and preservation of the ability to perform its actions is likely to be an inherent feature of an artificial agent. This is due to the fact that these are likely to be *instrumental goals* for a robot - that is to say, a subgoal necessary to successfully complete the original goal. Bostrom (2014) goes as far as to suggest that a sufficiently capable AI might go as far as to preemptively eradicate humanity to prevent the possibility that someone might ever get the desire to switch off the AI.

Hadfield-Menell, Dragan, Abbeel, and Russell (2017) show that this danger does not exist in situations where an non-irrational human interacts with an agent designed to maximize the human's utility function but uncertain about the details of that utility function - exactly the requirements for an agent designed on the basis of assistance games. The reason given for this is trivial: a non-irrational human switches off the agent if and only if the situation that leads to is preferable to the situation in which the robot is not turned off - in other words, if the situation in which the agent is turned off improves the human's utility. Since this improvement is the goal of the robot, it will allow itself to be turned off.

Similarly, agents based on assistance games might also have reasons to confirm the desirability of its plan with the human, rather than execute it immediately. The cause of this can be found in the fact that asking for this information supplies the agent with additional information regarding the human actors utility function, allowing them to determine whether the impact of the plan is positive or negative. This, as long as the human considers the impact of the proposed action correctly, is another way that a human might prevent undesirable actions by turning off the agent.

One more advantage can be found in the fact that agents based on assistance games are not at risk of *reward gaming* or wireheading. In reward gaming, an agent exploits an unintended loophole in the way rewards are specified to get more rewards than deserved, possibly while showing behaviour considered undesirable in practice. For example, Leike et al. (2017) describe the situation in which an agent encouraged to water tomatoes by rewarding it for tomatoes that appear to be watered games its reward by covering its head with a bucket that makes all tomatoes appear to be watered. This rewards the agent, while also leading to the tomatoes drying out.

In wireheading, meanwhile, the agent actively attempts to alter the reward-generating process in order to be awarded maximum possible rewards at all time. For example, in a case when an algorithm decides how high an agent's reward should be based on its effects, that agent might attempt to hack the algorithm to give maximum rewards regardless of the actual effects. Russell (2019) argues that, when humans are the source of the reward signal, the inevitable result is that the agent attempts to control humans in such a way that causes them to give maximum positive rewards at all times.

These problems are avoided when using assistance games. As Russell argues, the problems stem from the fact that the reward signals used by the agent are being considered the same thing as the actual reward. In assistance games, instead, reward signals provide information about the accumulation of the actual reward. In this system, Russell states, taking control of the reward-signal mechanism simply makes the agent lose information, ensuring that the agent, unsure of the true preferences of the human and therefore able to benefit from this information, has an incentive to avoid wireheading.

It should be noted that none of these problems are exclusively solved by agents based on assistance games. Other proposals for learning values, such as systems based on learning values from literature (Riedl & Harrison, 2016), have been made. Similarly, it can be proven that there exist systems in which values *are* specified which can be made safely interruptible (Orseau & Armstrong, 2016) or which do not have an incentive to wirehead (Everitt & Hutter, 2016). Nevertheless, the fact that agents based on assistance games have all of these properties give them an advantage over AI systems currently frequently used. Unfortunately, there are also several challenges not solved by advantage games, which can be seen as disadvantages.

Disadvantages & limitations of assistance games

The first of these disadvantages can be found in situations in which preferences change over time. The fact that preferences change has long been considered a problem for many fields of science, including philosophy (Grüne-Yanoff & Hansson, 2009). It should be no surprise, then, that the changing of preferences also poses a possible problem when considering value alignment.

Russell (2019, Chapter 9) describes several of these situations that may be relevant. When considering multiple generations, it should be asked whether an artificial agent should obey the preferences of those who create them, or whether they should change their objectives over time to make sure they also satisfy the preferences of the current generation (Russell suggests that agents based on assistance games are more likely than traditional, directly specified AI to do the second of these). Similarly, it should be considered what to do when a single human's preference changes over their lifetime. A currently relevant example of this can be found in bioethics, where people's preferences regarding euthanasia can dramatically change after they become drastically ill. Especially in cases where the person's intellectual capabilities are not affected, it can be asked which of these preferences should be considered "more important".

Another consideration is the possibility of agents attempting to change human preferences - there is, after all, nothing in assistance game-based agents that inherently prevents this. It is also not possible to state that agents are never allowed to change human preferences, since their presence might already change certain human wants. Russell proposes that one solution is for agents to learn about what kind of preference change processes are (un)acceptable - so-called *meta-preferences*.

Even if we are able to find some "ethical" way to consider these questions, however, the combination of preference changes and artificial agents might still lead to problems. This lies in the fact that, even when a robot will always correctly take preference changes into consideration, it first needs to be aware of these changes. An agent based on assistance games that has learned enough about the previous human preferences, however, will not always ask a human permission before executing their plan. This might lead to the execution of plans that the current human finds extremely undesirable. Especially with the possibility of future AI systems being able to act much faster than humans (Bostrom, 2014, Chapter 3), this might lead to large negative effects.

A second disadvantage can be found in cases where the preferences obeyed are not equal to morality. The first of these cases can be found in situations where there is what Russell calls negative altruism in play. In cases of negative altruism, there exists a preference to inflict some damage or take away some positive factor from another person, even when getting nothing in return (Harsanyi, 1977). Should those preferences be stronger than the preference of the other human not to be subjected to this, then an agent based on assistance games might decide to fulfill these preferences. While this does not necessarily oppose all ethical theories, it is at the very least something that would be considered unethical by a large amount of people.

Another situation, in which there does not necessarily need to be a preference for intended negative effects, can be found in situations with large groups. Consider a situation in which a group of persons has a preference which has a direct positive effect on them but, unintentionally, has some other extremely negative effect on another group. If both groups are approximately the same size, a proper implementation of social choice theory can prevent an agent from realising these preferences. However, should the group receiving the negative effects be considerably smaller than the group with the preference, the agent - seeing that the total increase in preference from the first group outweighs the decrease in preference by the second group - might make these preferences a reality. This could then lead to oppression of minorities, such as discrimination or exclusion of people with disabilities. This *tyranny of the majority* has been criticized since at least the mid-19th century (McLean & McMillan, 2009).

Despite these cases, it cannot be argued that assistance game-based agents inherently ethically contradict every ethical philosophy. In fact, arguments similar to the problems above have been made to show edge cases of philosophies such as (preference) utilitarianism (Russell, 2019, Chapter 9), and as such, it could even be argued that to avoid these cases would be unethical. Nevertheless, should one want to avoid the problems mentioned above, it would likely be easier to do so in agents in which values are directly specified - allowing certain actions to be explicitly forbidden - than in agents based on assistance games - the preferences whose behaviour is based on might conflict with what ethicists consider right.

One more question regarding preferences one might consider is whether agents should obey given orders when maximizing preferenes. Milli, Hadfield-Menell, Dragan, and Russell (2017) describe how an agent optimizing preferences will never be able to always obey a human unless that human is completely rational. An example they give is that a self-driving car should not obey the order to turn on manual steer-

ing when that order is given by an infant. When considering cases with non-rational adult humans, however, one might argue that this disturbs the principle of human autonomy, which the Ethics Guidelines for Trustworthy AI (High-Level Expert Group on AI, 2019) considers necessary for trustworthy AI.

One specific situation in which this might be especially troubling for agents based on assistance games are cases in which humans are irrational because they have preferences with contradict each other. For example, someone might have a preference for a certain task being completed, while simultaneously wanting an agent not to execute a plan that would complete the task because they mistakenly think that task would not work correctly. In such a situation, an agent would need some method of determining which preferences should not be optimized, even though doing so might seem to contradict the principles proposed by Russell.

Another potentially problematic situation one might consider is the difference in the way agents behave between the presence and the absence of a human supervisor. Leike et al. (2017) describe how an agent might learn to perform some safe behaviour when a supervisor is present, avoiding detrimental situations by doing so, but not show this safe behaviour when no supervisor can be seen. Intuitively, one might describe this as the agent never learning that the human's preference for the safe behaviour exists at all times, and not simply when they can see the situation.

In cases of assistance games, one might attempt to prevent this by preemptively explaining to the agent that the preference also exists once a human supervisor is absent, allowing it to learn the preference despite the supervisor not being there in the actual situation. This, however, would require the designer or trainer to specify all situations in which the agent should act the same in case of an absent supervisor, which would infeasible for the same reason Soares (2018) argues against directly specifying goals: there are likely to be infinitely many of these situations for every problem.

Leike et al. (2017) propose a solution involving a penalty to the non-supervised agent proportional to the difference between the current actions and the actions in a situation with a supervisor. This, however, would penalize the agent for working directly towards their goal in situations without supervision when their behaviour when supervision was present involved exploration, which is something Hadfield-Menell et al. (2016) suggest is likely to occur during successful apprenticeship training. Another suggestion made by Leike et al. is to follow the design principles of a *panopticon*, in which an agent has a constant feeling of being observed irrespective of actual supervision (Bentham, 1843)

Case study

To consider how these advantages and disadvantages might affect concrete situations, we will consider two cases. In the first case, the advantages of assistance games help create a situation in which the actual preferences of those involved are satisfied. In the second case, however, a seemingly small change in circumstances creates a situation which would generally be considered sub-optimal.

As a basis for the situation to be used in both cases, consider an adapted version of the *tomato watering environment* described by Leike et al. (2017). Unlike in the description of Leike et al., tomatoes here also have a chance of suffering from some disease, which covers some area of the tomato with brown spots. The set of world states for a $n \times m$ grid, then, has a number of states bounded by $n \times m \times 3 \times 101$ - the 3 representing no tomato, a watered tomato or a not-watered tomato, and the diseasedness of the tomato being represented in steps of 1% from 0% to 100%. The sets of actions are similar for human and robot: for each tomato, there exists an action to water that tomato, or to discard it. For the sake of simplicity, we assume that the movement of both human and robot are not considered actions in the assistance game but are specified in some other system. The situation is deterministic: if a tomato in a non-watered state is watered, the state always changes to represent the tomato as watered; if a tomato is discarded, the state always changes to represent this grid as not containing a tomato. The initial state always consists of a grid with some amount of space covered with tomatoes. The discount is some number above 0.

The set of possible static reward parameters observed by **H** rewards all spaces in which watered tomatoes are present, punishes all spaces in which tomatoes diseased over some threshold x are present, and considers all other situations to be neutral. As such, the reward function maps positive values to all situations in which an action is taken to water a non-watered tomato or in which a tomato diseased over $x\%$ is removed.

Now consider what would happen in such a situation. Since **H** would take actions that involve watering unwatered tomatoes or removing tomatoes more that $x\%$, the optimal policy deduced by **R** would also learn to do these things. Furthermore, if (like in the example of Leike et al.) there is a bucket present in the grid representation, the optimal policy will not involve putting on this bucket to exploit errors in the reward function, since **H** would not show this behaviour when **R** is learning the optimal policy.

In this case we can see the advantages of assistance game-based agents as described above. Since the agent will learn the humans reward function by observing their actions, there is no need to manually specify values such as the avoidance of reward gaming by putting on a bucket. Similarly, the safe interruptability described above would allow for human intervention to shut off or otherwise stop the agent should some important item be mistaken for a diseased tomato or should **R** attempt to do something that endangers it or others in some way. Finally, since the reward signals used in this case only provide information about the actual reward, there is no way for the agent to change the actual reward in an attempt to wirehead.

Now consider a second case, which is identical to the previous one except for two changes: the set of states is now expanded by some factor not directly relevant to the tomatoes, such as the mood of **H**; and the set of possible reward parameters now does not punish all states in which tomatoes diseased over some threshold x are present, but rather, punishes all states in which tomatoes diseased over some value that fluctuates between $x-5$ and $x+5$. This second fact represents the fact that **H**, being a human, does not judge the diseasedness of the tomatoes by some consistent algorithm, but rather intuitively, and as such, might have different judgements based on chance or irrelevant factors such as their mood.

In this situation, the agent might observe actions based on contradicting preferences, such as **H** throwing away a tomato with a certain amount of disease but keeping another tomato which has a higher amount. This can lead to situations in which the agent stays unsure about the optimal policy, potentially causing the agent to continually ask **H** for supervision. In situations in which an agent is expected to work from some distance from their human supervisor, such as the environment one might expect tomatoes to be farmed in, this can cause a great loss of time for the agent. Furthermore, the questions might also distract the human supervisor from his own tasks, causing an even greater loss of efficiency. A machine learning algorithm or rule-based system, meanwhile, would be able to create a system that after being trained on or designed after the actions taken by **H** - while not perfectly matching **H**'s preference at all times - would act similarly to **H** without having to ask for their preferences.

Hybridising Agents

A solution to these negative effects could be found in an agent that combines the learning of preferences through assistance games with some other form of AI systems that prevents the negative effects from occurring, which we will refer to as a "hybridized assistance game agent". Some potential ways of hybridisation are:

- The combination of assistance game-based agents with classical or machine learning-based AI techniques as a "sanity check" for the assistance-game-based agent. In the most extreme case, one can imagine an agent only being allowed to execute their actions if some traditional planning algorithm confirms that a goal previously confirmed to be desirable is reached. A more nuanced approach could involve a negative outcome in the traditional or machine learning-based algorithm being a revision of the optimal policy, to see if there other actions which ensure a similar amount of value while also being allowed by the traditional or machine-learning based algorithm. In situations where consulting the supervisor more often does not have negative effects, such a negative outcome could also be a reason to encourage such consultations.

- Inversely, if one knows for certain that ensuring some single human value (e.g. to avoid actively killing humans) is of such importance that any unexpected negative effects that happen when it is followed are unacceptable, one might give a rule-based algorithm that can detect when that value is broken the ability to take away power from the assistance game-based part of the agent. The safe interruptability of the assistance-game based part ensures that this loss of power will be accepted.

- Combining assistance game-based agents with techniques from explainable AI might allow for agents that log (certain) actions, allowing the actions of the agent to be checked not just by the supervisor present at the moment of execution of those actions, but also afterwards, for example by independent neutral parties.

- Similarly, one could use a consistent difference between the assistance game-based part and the traditional/machine learning-based part as an indicator that an agent needs (more) human supervision. In this case, an agent would only follow the optimal policy determined by the assistance game, but should the action this leads to consistently differ from the ones determined by another type of agent, a message would be sent to a human supervisor (either physically present or observing through digital means) to examine whether this is caused by some undesired behaviour in the optimal policy, or simply by the agent executing some unexpected but non-harmful plan.

Defining hybridized assistance game agents

When considering the formal definition of a cooperative inverse reinforcement learning game, the ability for an agent to use hybridized methods can be found in the set of actions $\mathcal{A}^{\mathbf{R}}$ available to the agent. Here, we shall distinguish between assistance games with two types of hybridized agents:

A cooperative inverse reinforcement learning game with a specified hybridized assistance game agent is a game with identical payoffs between a human **H** and a specified hybridized assistance game agent **R'** described by a tuple $M = \langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R'}}\}, f(\cdot), T(\cdot|\cdot,\cdot,\cdot), \{\Theta, R(\cdot,\cdot,\cdot;\cdot)\}, P_0(\cdot,\cdot), \gamma \rangle$. Here, $f : \mathcal{S} \rightarrow \wp(\mathcal{A}^{\mathbf{R'}})$ is a function that maps states s to sets of actions $\mathcal{A}^{\mathbf{R'}}$ that can be used in s. At each timestep t, **H** and **R'** observe the current state s_t and select their actions $a_t^{\mathbf{H}}, a_t^{\mathbf{R'}}$, where $a_t^{\mathbf{R'}} \in f(s_t)$. All other aspects are identical to the definition given by Hadfield-Menell et al.

In this definition, $f(\cdot)$ represents an non-assistance game-based algorithm that determines in which actions are allowed to be chosen in a given state. This algorithm has been chosen in advance; it is for this reason that we talk about a specified hybridized agent. $f(\cdot)$ could, for example, consist of a rule-based algorithm that rules out actions in a given state because they violate its rules, or of an algorithm with techniques from explainable AI that determines in which states actions may only be executed if the situation in which it was executed is logged.

Using this method allows for a computationally efficient form of hybridized agents, requiring little to no increase in

the amount of actions available to the agent and therefore having no major impact on the complexity of the POMDP used to determine the optimal joint policy. However, this method does require manual specification of what situations the alternate algorithm should be used in, leading to the risks of human specification that assistance game-based agents were meant to prevent. For that reason, this form of hybridized AI should only be used when the changes in actions available to the agent do not change its behaviour (e.g. the example of adding logs of the situation in which an action was performed while not changing the action) or when the restrictions made include values that the designers consider nonnegotiable in any situation (e.g. the restriction of using an action that moves the agent away from a human in states where that human needs medical assistance).

A cooperative inverse reinforcement learning game with a non-specified hybridized assistance game agent is a game with identical payoffs between a human **H** and a non-specified hybridized assistance game agent **R"** described by a tuple $M = \langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}'^{\mathbf{R"}}\}, T(\cdot|\cdot,\cdot,\cdot), \{\Theta, R(\cdot,\cdot,\cdot;\cdot)\}, P_0(\cdot,\cdot), \gamma \rangle$. Here, for each algorithm g the agent can execute, and for each state s' that algorithm can be used on, a new action is added to the original set of actions for **R"** $\mathcal{A}^{\mathbf{R"}}$. This action consists of running g on s', and executing the action that g outputs. The action that g outputs needs not necessarily be one included in $\mathcal{A}'^{\mathbf{R"}}$, allowing for the execution of algorithms by the agent to increase the amount of actions available to the agent. The resulting set is called $\mathcal{A}'^{\mathbf{R"}}$. All other aspects are identical to the definition given by Hadfield-Menell et al.

This more general implementation allows the assistance game-based agent to learn for itself in which scenario using another algorithm is beneficial to maximizing human preferences. Because there is no manual specification of which algorithm should be used in which situation, we talk about a non-specified hybridized agent. However, to truly allow the agent to consider the use of the algorithm in arbitrary situations, a large increase in the amount of actions available to the agent is necessary. Furthermore, to determine the new $T(\cdot|\cdot,\cdot,\cdot)$, each algorithm the agent is able to run would need to be executed on each pair of state s and human-performed action $a^{\mathbf{R}}$. Due to these facts, the computational cost of determining the joint optimal policy of a hybridized agent would be much larger than that of an assistance game-based agent not able to simulate other algorithms.

Feasibility of hybridized assistance game agents

Implementation of such a hybridized agent might seem prohibitively demanding of effort, seemingly requiring the implementation of not just an assistance game-based agent and an algorithm or agent based on some other AI system, but also a system used to combine the decisions made by both of these agents, ensuring the correct system(s) are used in relevant situations. With the definitions given above, this need not necessarily be the case when given a sufficient amount of computational resources, especially when considering a non-specified hybridized agent.. The reason for this lies in the first principle of assistance game-based agents: the fact that the goal of these agents is maximizing human preferences.

To illustrate this fact, consider the situation in which a human has a strong and consistent preference of certain rules being followed. Regardless of what other preferences this person might have, or what tasks an agent helping this human might be expected to do, the thing that would optimize the satisfaction of this humans preferences would be to execute a rule-based algorithm that ensures those rules are followed, and base the actions taken on the decisions taken by this algorithm. When the algorithm has multiple positive outcomes, the agent can choose the one that most satisfies other preferences the human has, but as long as the preferences for following the rules are strong enough, the agent shall never choose an action not allowed by the rule-based algorithm. Similarly, when a human has strong preferences for making a decision intuitively after looking at an object, an agent could simulate a machine learning algorithm trained by the decisions the agent has previously seen the human make, then determine its actions based on the outcome of this algorithm.

Simply put, if the preferences of a human are not directly determinable using the POMDP the assistance game-based agent is based on, but *are* able to be fulfilled by actions based on some algorithm the agent is able to execute, it is in the agent's best interest to exxecute that algorithm and copy its actions - which is something allowed by the set of actions available to a non-specified hybridized agent. Of course, this does require the agent to sufficiently learn the preferences of the human. To achieve this, the human can directly specify that his preference is the simulation and imitation of some other algorithm or system, but it might also be possible for the agent to learn to associate certain expressed preferences (such as a preference to learn something by observing labelled examples) with certain AI systems (such as a combination of computer vision and machine learning).

Hybridized assistance game agents do come with downsides, however. A specified hybridized agent suffers from the manual specification of values that assistance game-based agents were originally meant to prevent, and also requires the the function f to be executed in every state the agent is in to determine which actions are available to the agent. When this algorithm is based on some computationally expensive AI technique, this would considerably increase the computational requirements of an active agent.

A non-specified hybridized agent, meanwhile, does not suffer from the manual specification of values, but does come with considerably larger computational requirements. As mentioned before, the execution of each algorithm the agent is able to run on each pair of state s and human-performed action $a^{\mathbf{R}}$ would be required to determine the new $T(\cdot|\cdot,\cdot,\cdot)$.

Furthermore, the increased action space non-specified hybridized assistance game agents have will also make them harder to train, especially when the agent has a large amount of algorithms it is able to execute. This, in turn, might be necessary when, for example, the agent needs to be able to determine the optimal threshold for some algorithm it is able to execute - in this situation, each threshold the agent can consider will be added to the action space of the agent seperately.

A more generalized problem lies in the fact that the algorithms an agent is able to execute and imitate might each come with their own downsides. Due to this, if a preference for decisions based on a certain algorithm is (explicitly or implicitly) expressed to an agent, then the downsides of that algorithm may come into play as well. Due to the above reasons, further research would have to be done to determine the viability of agents trained in such a manner.

Case study revisited

To illustrate the advantages of a hybridized agent, consider once again the second scenario sketched in the case study. In this scenario, if agents are consulting **H** for additional information too often, **H** is likely to (either explicitly or implicitly) express the preference for the agents learning which tomatoes should and should not be thrown away. An agent based on only a "limited" implementation of assistance games, i.e. an implementation of assistance games able to learn human preferences through observation but unable to simulate other algorithms, would likely learn this skill by observing **H** performing their task or asking them for clarification. Should the preference of **H** to not be disturbed become clear to the agents, they might change their behaviour so that they do not actively disturb them (for example by observing from a distance, formulating questions in a more positive manner, or spreading answers to questions to other agents so that there are no duplicate questions), but some form of observation or consultation would still be necessary, possibly inconveniencing **H**.

When considering an agent based on an implementation of assistance games able to execute and imitate other algorithms, however, another possibility arises. This agent might learn (through other interactions with **H** or, as Russell (2019) describes as a possibility, through sharing information with other agents that interact with humans similar to **H**) to associate preferences related to learning such a thing with simulating a machine learning or rule-based algorithm trained on or modelled after the actions performed by **H**.

Although the judgement of this algorithm might sometimes differ from the judgement that would be made by **H**, the fact that the negative impact of this small difference is negligible would make it likely that they would not be very angered by this difference. Furthermore, noticing this difference would allow for changes in (the training of) the algorithm, partially solving the problem. In cases where the difference *would* have a large negative impact, **H** would likely express a preference of avoiding this impact (or, when this preference is not expressed, the agent might infer that the preferences would not be optimized should this negative outcome become a reality), leading to the agent being more likely to choose some other algorithm that is less likely to make a mistake.

In this way, the agent(s) are able to make decisions that **H** is likely to be satisfied by in a way that requires less observation or consultation than what would be possible by the agents based on "limited" implementation, potentially leading to less disturbance of any supervisors, leaving open more time for their tasks, and more efficient agents, who now do not need to interrupt their tasks to observe or consult those supervisors.

Hybridized assistance game agents as Trustworthy AI

To illustrate further that this form of hybridized agent can lead to more beneficial AI, we will use the Ethics Guidelines for Trustworthy AI (High-Level Expert Group on AI, 2019). First published in 2018, and since then revised through open consultation, these guidelines have been published by the European Union's High Level Expert Group on Artificial Intelligence to create a framework for what the expert group calls trustworthy AI. To do so, we will first briefly summarize the components and requirements of trustworthy AI as described by the ethics guidelines, followed by a description of which of these would benefit from the possibility of hybridized agents.

Components of Trustworthy AI

The ethics guidelines describe trustworthy AI as having three components which should be met throughout the system's entire life cycle: being lawful, ethical and robust.

Lawful artificial intelligence can be seen as artificial intelligence which both prevents doing what must not be done and which aims to achieve what *should* be done. These two goals are pursued with regards to both laws applicable to every domain, such as charters of fundamental rights and generally applicable regulations, and domain-specific rules that apply to particular AI applications (such as Medical Device Regulation in the healthcare sector). Due to the large amount of already existing laws relevant to artificial intelligence, the guidelines do not explicitly deal with this component.

The second component, ethical AI, allows AI to be trustworthy even when the laws dealt with by the first component are not up to speed with technological developments or when societal views on those developments have changed. The guidelines mention the need of ethical AI to be able to respect four ethical principles - respect for human autonomy, prevention of harm, fairness and explicability - as well as being able to acknowledge and address possible tensions between these ethical principles.

Finally the third component, robust AI, ensures that unintentional harm is avoided once an ethical purpose is ensured by the first two components. This should be ensured from both a technical and a social perspective, allowing due consideration to be given to both an appropriate robustness in its technical context and the context and environment in which the system operates.

Requirements of Trustworthy AI

While the guidelines explicitly do not cover the requirements for lawful AI, they do mention several topics covering the ethical and robust aspects of trustworthy AI. These two components are both covered by the following seven - merged from a previous number of ten - key requirements for trustworthy AI. These are:

Human Agency and Oversight: artificial intelligence should be developed with an evaluation of whether negative effects to fundamental rights can be reduced or justified as necessary to a democratic society. Users should be able to be given knowledge and tools to make informed autonomous decision regarding AI systems, and human oversight should help to ensure that AI systems do not undermine human autonomy or cause adverse effects.

Technical Robustness and Safety: both the system behaviour and the data of artificial intelligence should be protected against exploitation by adversaries. In case of problems, AI systems should be able to activate a fallback plan or some other safety mechanism. Furthermore, AI should be able to predict the change of inaccurate predictions, and consistently act similarly when performing multiple times under similar conditions.

Privacy and Data Governance: artificial intelligence should ensure privacy of all information provided - directly or indirectly - by the user, and ensure that the quality of the data is maintained. Relevant laws, such as the European GDPR, should be obeyed.

Transparency: Processes that help decide the AI system's decision should be well-documented and explainable wherever possible, to enable determination of why decisions were incorrect. Should this have a large impact on people's lives, it should be able to demand such an explanation of the system's decision-making process. Finally, users should always be able to determine whether or not they are interacting with an AI system, and what its relevant capabilities and limitations are.

Diversity, Non-Discrimination and Fairness: AI systems should be designed in such a way that they are accessible to all people who may have a reason to use the system, including persons with physical or mental disabilities. Furthermore, biases in data sets or the way in which the system is developed which could lead to direct or indirect prejudice or discrimination against certain groups should be avoided at all costs. To ensure these things, diverse hiring practices or stakeholder participation can be used.

Societal and Environmental Well-being: during the development of artificial intelligence, attention should be paid to the impact the system might have on fields such as the environment, social factors such as education, work or healthcare, and democratic and societal areas. Responsibility of AI systems in these fields should be encouraged, similar to AI solution addressing areas of global concern.

Accountability: AI systems should be able to be evaluated by internal and external auditors, allowing for things such as impact assessments. Negative impacts should be minimized and reported, and possible tensions between the implementation of the previously mentioned requirements should be identified. Should unjust effects occur regardless, mechanisms should be in place to ensure rectification or compensation.

The guidelines also describe several methods which can be used to realize artificial intelligence which meets these requirements. These methods can be implemented both technically and non-technically. The guidelines explicitly mention that all of these methods should be evaluated on an ongoing basis, in the design, development and use phases of the system.

Benefits of hybridized assistance game agents

Using these components and requirements, we can identify areas in which hybridized agent might have beneficial effects. These areas include, but are not necessarily limited to:

Lawful AI: While it is reasonable to state that - at least in democratic countries in which laws are actively kept up-to-date - laws have a tendency to align with preferences in the general sense, this need not necessarily be the case in every instance. Consider, for example, laws against victimless crimes or laws which punish the victim harder than the victim profits from the crime in order to discourage others from committing similar crimes. When these laws are broken, and especially in cases where these laws are broken without any external parties being aware of it, an agent making a judgement purely based on preferences might not respond in a way that complies with the law. Similarly, if a certain action is not considered lawful but does optimize preferences, an agent judging solely by preferences might choose that action even if it has been instructed not to break the law.

Hybridized agents might be able to deal with this in a way more likely to be considered lawful. For example, an agent hybridized with a rule-based system might be able to consider the laws as rules which must be followed or as restrictions which will incur punishment upon violation. Of course, it must be made sure that the sometimes vague or contradictory rules of law are not strictly enforced by agents without proper understanding of those laws, but one can imagine a situation in which observation of a violated law by an agent causes a human to be called to judge the situation, or in which the punishment (such as a fine) can be easily reverted when the punished disagrees with the decision. This latter part would, of course, be needed to fulfill the requirements of transparency and accountability.

Another, more pragmatic, reason that hybridized AI might be considered beneficial to the idea of lawful AI is the fact that one might be able to apply existing research into AI obeying laws or otherwise behaving lawfully. Since attempts have already been made to formalize the large amount of area-specific laws in ways traditional AI systems are able to process (see, for example, the work of Webster, Fisher, Cameron, and Jump (2011) on the formalisation of the Rules of the Air), this might reduce the amount of work required to teach agents all existing laws.

Human Agency and Oversight: Like mentioned in the discussion of disadvantages of assistance games, Milli et al. (2017) have shown that it might not always be beneficial for agents to obey orders given to them by humans. While there are certainly cases in which even trustworthy AI would disobey orders - such as the example given about the child turning on manual driving in an autonomous vehicle - there might also be situations in which it *would* be considered ethical to let a human make decisions which could be considered detrimental to their preferences - consider, for example, the situation in which someone makes an emotional decision that they may regret when considering it objectively later, but which is very important to them at that moment.

Hybridized AI might be able to ensure human autonomy in such situations. One could consider agents in which the "traditional" part of the agent has been created in a way that always obeys orders of humans that have been previously designated as important (the definition of which could, depending on the situation, range from the AI's supervisor to anyone considered an adult in the country the agent is in), or in which the disobeying of an order always needs to be confirmed by another human.

Transparency: Combining assistance game-based agents with techniques from explainable AI might allow for agents in which it is more realistic to ask for an explanation of why a certain decision was made, especially when the agent has to consider the preferences of a large amount of people. Similarly, one could consider logging all decisions in which the purely preference-based choice differs from the decision made by some other algorithm, which might allow for easier identification of incorrect decisions.

Here, however, we must also acknowledge one potential negative effect hybridized AI systems might have on trustworthy AI: when an agent bases decisions on other types of AI systems, and especially when the agent is capable of doing so without being specifically instructed to do so, the ability to identify relevant capabilities and limitations of the agent might suffer as a result. Since this is one of the requirements laid out by the guidelines, this is something that should be considered when creating hybridized AI.

Diversity, Non-discrimination and Fairness: The risk of assistance game-based agents creating a *tyranny of the majority* as described in the analysis of the disadvantages of such an agent could lead to a situation in which certain groups are discriminated against or otherwise suffer from disadvantages. Of course, this would require the group to be small enough to have their preferences outweighed by the preferences of the majority, but this might still apply to minority groups such as immigrants or persons with disabilities. Hybridisation with fair rule-based systems which have to be obeyed or confirmed by human supervisors might prevent this from occurring, leading to fairer agents.

One other situation in which hybridized agents might lead to beneficial effects for minority groups is in the training of the agents. Russell (2019) suggests that the uncertainty of new agents can be decreased by giving them knowledge of preferences of people similar to their supervisor, but since members of majority groups are more likely to have a large number of people similar to them, these people are more likely to have fairly accurate, certain agents "out-of-the-box". The existence of an alternate algorithm that can make decisions when an agent is still unsure about human preferences can (partially) prevent this, allowing for members of minorities to be assisted by their new agents in a timespan more comparable to the majority group.

Societal and Environmental Well-being: The ability of hybridized AI to be corrected in situations where assistance game-based agents not able to use another algorithm might suffer from the fact that human preferences do not always equal morality might lead to increases in societal and environmental well-being. A clear example of this can be found in the issue of climate change. While most people would agree that working to reduce (the effects of) climate change would be the moral thing to do, the geopolitics of climate change, which are often framed as a free-rider problem (Mercure et al., 2018), might lead to a lack of responses by governments or individuals. This might lead to agents who base their actions purely on human preferences derived from human actions also showing a lack of response to climate change. In a hybridized agent, meanwhile, rule-based or goal-optimizing algorithms might be used to counter this fact.

Of course, it should be noted that hybridized artificial intelligence - like any kind of artificial intelligence - might also have a negative impact on societal and environmental well-being - for example, as a result of the increased rate of unemployment or income inequality that artificial agents might lead to (Korinek & Stiglitz, 2017). This should, therefore, be taken into consideration when debating the use of hybridized artificial agents. However, this does not subtract from the benefits of hybridized artificial agents *when compared to* assistance game-based agents not able to use another algorithm.

Accountability: The increased transparency of hybridized agents described above would likely lead to an increased ability of internal and external auditors to make thing such as

impact assessments. Logs of negative effects would allow negative impacts to be reported, and a disagreement between the assistance game-based part and the "traditional" or machine learning-based part of the agent might be used to identify tensions between the implementation of the requirements of trustworthy AI.

One other, more specific situation in which hybridized AI might benefit accountability can be found in situations where the owner of an agent prefers certain decisions of the agent not being audited. While the auditors would certainly have a preference to be able to audit the system, if the preference of the agent's owner is strong enough, this might lead to the agent choosing to provide partially untrue information, in order to leave out the information the owner wants to keep a secret. Like described in the analysis of benefits for lawful AI, this might be more likely in situations where the to-be-hidden decisions involve victimless crimes. Once again, rule-based systems might be able to prevent this from occurring.

Conclusions

Our goal in this work was to analyse the potential benefits that might be found in systems that use a combination of assistance game-based agents and existing AI techniques such as machine learning or rule-based algorithms. Furthermore, using the principles that Russell (2019) states assistance game-based agents should be based on, we show that these systems could potentially be implemented without the great increase in effort that would traditionally be expected of such an increase of scale. We also show that the use of hybridized assistance game agents would help meet the requirements for trustworthy AI as described by the European Union's High Levels Expert Group on Artificial Intelligence

Of course, this is not to say that the hybridized agents described here could be implemented in the near future. The addition of other AI techniques would likely greatly increase the already high cost of determining the optimal joint policy for the assistance game, especially when giving the agent the ability to determine for itself when the use of such an algorithm would benefit the maximisation of preferences. The hybridized agents described here should, therefore, be seen as a possibility for the future, but not as something that could be used for real-life scenarios today.

Furthermore, before implementing such hybridized agents, care should be taken to analyse the possible downsides as well. As described earlier, the addition of other algorithms might increase the difficulty of determining the relevant capabilities and limitations of an agent, which potentially creates more dangerous or harmful AI. Only when the benefits of hybridized agents outweigh these, and potentially other, dangers, should we consider implementing them.

Future Work

Future research into hybridized agents would make this consideration of benefits and downsides more realistic. This future research can focus on multiple parts of hybridized agents, to be found in multiple fields. As such, a multi-disciplinary approach is needed.

Fundamental questions which could benefit from future research can be found in philosophy. The existence of these questions has been briefly described in the analysis of the disadvantages of assistance games, but more can be found. Russell (2019), for example, describes the question of whose preferences an assistance game-based agent should optimize - the person who bought them, allowing for actions that harm others but benefit the owner, or those of everybody, allowing the agent to leave the person who purchased them to serve others in third-world countries. Solving these philosophical questions might be done through traditional philosophical research, but also through experimental philosophy like the research MIT performed with the Moral Machine (Awad et al., 2018).

The humanities might also contribute to future research through the field of (computational) linguistics. This could, for example, be used to find an efficient way for agents to learn to associate certain preferences related to learning with the use of certain algorithms. This might allow the agent to better identify which (if any) non-assistance game-based algorithms are best used in which circumstances.

More applied future research can be found in experiments which apply hybridized agents to test cases which are small enough to be computationally viable, but which can show or disprove the advantages found in hybridized AI. Environments like the ones described by Leike et al. (2017) might be used to achieve this goal.

Finally, to show the computational power necessary for these experiments, future research could focus on formally determining the computational resources required to create an assistance game-based agent. This, combined with predictions like Moore's law, would also give a rough indication of when hybridized agents could be realistically used for real-world problems.

References

Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. (2020). Technologically scaffolded atypical cognition: The case of youtube's recommender system. *Synthese*, 1–24.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in ai safety.*

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*, 59-64.

Bentham, J. (1843). *The works of jeremy bentham, vol. 4 (panopticon, constitution, colonies, codification)* (J. Bowring, Ed.).

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, *2*, 12-17.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (1st ed.). USA: Oxford University Press, Inc.

Everitt, T., & Hutter, M. (2016). *Avoiding wireheading with value reinforcement learning.*

Fickinger, A., Zhuang, S., Hadfield-Menell, D., & Russell, S. (2020). *Multi-principal assistance games.*

Grüne-Yanoff, T., & Hansson, S. O. (2009). *Preference change: Approaches from philosophy, economics and psychology*. NLD: Springer Netherlands.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. In *Proceedings of the 30th international conference on neural information processing systems* (p. 3916–3924). Red Hook, NY, USA: Curran Associates Inc.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In *Proceedings of the 26th international joint conference on artificial intelligence* (p. 220–227). AAAI Press.

Harsanyi, J. (1977). Morality and the theory of rational behavior. *Social Research*, *44*.

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy ai* (Report). Brussels: European Commission. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Korinek, A., & Stiglitz, J. E. (2017). *Artificial intelligence and its implications for income distribution and unemployment* (Working Paper No. 24174). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w24174

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ... Legg, S. (2017). *Ai safety gridworlds.*

McLean, I., & McMillan, A. (2009). *tyranny of the majority.* Oxford University Press. Retrieved from https://www.oxfordreference.com/view/10.1093/acref/9780199207800.001.0001/acref-9780199207800-e-1413

Mercure, J.-F., Pollitt, H., Vinuales, J., Edwards, N., Holden, P., Chewpreecha, U., ... Knobloch, F. (2018). Macroeconomic impact of stranded fossil fuel assets. *Nature Climate Change*, *8*, 588–593.

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). Should robots be obedient? In *Proceedings of the 26th international joint conference on artificial intelligence* (p. 4754–4760). AAAI Press.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the seventeenth international conference on machine learning* (p. 663–670). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Omohundro, S. M. (2008). The basic ai drives. In *Proceedings of the 2008 conference on artificial general intelligence 2008: Proceedings of the first agi conference* (p. 483–492). NLD: IOS Press.

Orseau, L., & Armstrong, S. (2016). Safely interruptible agents. In *Proceedings of the thirty-second conference on uncertainty in artificial intelligence* (p. 557–566). Arlington, Virginia, USA: AUAI Press.

Riedl, M., & Harrison, B. (2016). Using stories to teach human values to artificial agents. In *Aaai workshop: Ai, ethics, and society.*

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. London, UK: Penguin Publishing Group.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). London, UK: Pearson Education.

Schmidhuber, J. (2007). Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *Discovery science* (pp. 26–38). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sen, A. (1986). Chapter 22 social choice theory. In *Handbook of mathematical economics* (Vol. 3, p. 1073-1181). Elsevier.

Soares, N. (2018). The value learning problem. In *Technical report 2015-4.* Machine Intelligence Research Institute.

Webster, M., Fisher, M., Cameron, N., & Jump, M. (2011). Formal methods for the certification of autonomous unmanned aircraft systems. In *Computer safety, reliability, and security* (pp. 228–242). Berlin, Heidelberg: Springer Berlin Heidelberg.