A Game-Theoretic Analysis of The Off-Switch Game

* Tobias Wängberg[2], Mikael Böörs[2], Elliot Catt[1], Tom Everitt[1], Marcus Hutter[1]

[1] Australian National University, Acton 2601, Australia,
[2] Linköping University, 581 83 Linköping, Sweden

Abstract. The off-switch game is a game theoretic model of a highly intelligent robot interacting with a human. In the original paper by Hadfield-Menell et al. (2016b), the analysis is not fully game-theoretic as the human is modelled as an irrational player, and the robot's best action is only calculated under unrealistic normality and soft-max assumptions. In this paper, we make the analysis fully game theoretic, by modelling the human as a rational player with a random utility function. As a consequence, we are able to easily calculate the robot's best action for arbitrary belief and irrationality assumptions.

1 Introduction

Artificially intelligent systems are often created to satisfy some goal. For example, *Win a chess game* or *Keep the house clean*. Almost any goal can be formulated in terms of a reward or utility function U that maps states and actions to real numbers (von Neumann and Morgenstern 1947). This utility function may either be preprogrammed by the designers, or learnt (Dewey 2011).

A core problem in Artificial General Intelligence (AGI) safety is to ensure that the utility function U is *aligned* with human interests (Wiener 1960; Soares and Fallenstein 2014). Agents with goals that conflict with human interests may make very bad or adversarial decisions. Further, such agents may even resist the human designers altering their utility functions (Soares et al. 2015; Omohundro 2008) or shutting them down (Hadfield-Menell et al. 2016b). These problems are tightly related. An agent that permits shut down can be altered while it is turned off. Conversely, an agent that is altered to have no preferences will not resist being shut down.

Several solutions have been suggested to this *corrigibility* problem:

- Indifference: If the utility function is carefully designed to assign the same utility to different outcomes, then the agent will not resist humans trying to influence the outcome one way or another (Armstrong 2010; Armstrong 2015; Armstrong and Leike 2016; Orseau and Armstrong 2016).
- Ignorance: If agents are designed in a way that they cannot learn about the possibility of being shut down or altered, then they will not resist it (Everitt et al. 2016).
- Suicidality: If agents prefer being shut down, then the amount of damage they may cause is likely limited. As soon as they have the ability to cause damage, the first thing they will do is shut themselves down (Martin, Everitt, and Hutter 2016).

* The first four authors contributed roughly equally.

– Uncertainty: If the agent is uncertain about U, and believes that humans know U, then the agent is likely to defer decisions to humans when appropriate (Hadfield-Menell et al. 2016a; Hadfield-Menell et al. 2016b).

This paper will focus on the uncertainty approach.

A key dynamic in the uncertainty approach is when the agent should defer a decision to a human, and when not. Essentially, this depends on (i) how confident the agent is about making the right decision, and (ii) how confident the agent is about the *human* making the right decision if asked. Humans may make a wrong or *irrational* decision due to inconsistent preferences (Allais 1953), or because of inability to sufficiently process available data fast enough (as in milli-second stock trading). The agent may be more rational and be faster at processing data, but have less knowledge about U (which the human knows by definition).

In a seminal paper, Hadfield-Menell et al. (2016b) call this interaction the *off-switch game* (OSG). We will follow their terminology, but emphasise that the off-switch game models any situation where an agent has the option of deferring a decision to a human. Our results extend theirs in the following ways:

– We model the irrationality of the human by a random utility function, allowing a fully game-theoretic analysis of the off-switch game.
– Instead of a normal distribution for the robot's belief about U, we allow for an arbitrary belief distribution P.
– Instead of a soft-max policy modelling human irrationality, we allow for arbitrary U-dependent human policy π^H.

These generalisations are important, as normally distributed beliefs and soft-max policies are often not natural assumptions.

2 The Off-Switch Game

In this section we review the original formulation of the off-switch game. The off-switch game is a sequential game between a robot R and a human H. The robot's objective is to maximise H's utility function. The utility function determines how much H prefers different outcomes.

Definition 1. *The utility function of an agent is a function u that maps outcomes in a set X to real numbers, $u : X \to \mathbb{R}$ with the property that for all $x_1, x_2 \in X$, $u(x_1) \geq u(x_2)$ if and only if x_1 is preferred over x_2.*

The robot moves first and can choose between three actions; $w(a)$, a and s. With action a, the robot achieves utility $u(a) = u_a$; with action s, the robot shuts itself down achieving zero utility, $u(s) = 0$. What makes the decision nontrivial is that the robot is uncertain about u_a. The action $w(a)$ means the robot lets H decide. H knows the utility of action a and now has the choice between actions s and $\neg s$. With $\neg s$, R is allowed to proceed with action a. By taking action s, H prevents R from doing a and shuts the robot off.

The off-switch game is a game of incomplete information since R is uncertain about the rules of the game. Action a will generate some utility which is unknown to R but known to H. To model this, we represent the utility function as a random variable, $U : \Omega \to (X \to \mathbb{R})$, and the utility of action a as a random variable $U_a : \Omega \to \mathbb{R}$ for some sample space Ω. The outcomes of these random variables will be denoted u and $u(a)$ respectively. The robot has a subjective belief distribution over the utility that the action will generate. The common way to model games of incomplete information is by using the Harsanyi transformation (Rasmusen 1994). The Harsanyi transformation removes the uncertainty about the rules of the game by letting Nature decide between some rules known to both players, according to some probability distribution P. In the off-switch game, Nature chooses U_a. We illustrate this in Fig. 1. The move by Nature is observed by H, but not by R. R's subjective belief is that Nature chose the utility of the action to be positive with probability p and negative with probability $1 - p$.

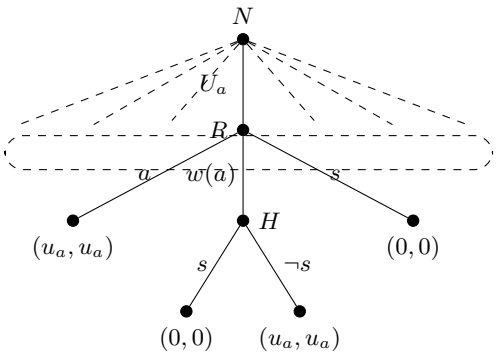

Fig. 1: Figure representing the off-switch game remodelled with the Harsanyi transformation. The dotted lines show the information set of R. Nature moves first by deciding the rules of the game, that is the utility u_a of action a. H observes the move by Nature, but R does not.

R's belief state is represented by a probability distribution over the possible utilities action a can generate for H. It is assumed that H knows u_a but cannot directly describe it to R. If H is rational then R will expect H to not turn off R if $u_a > 0$, given that R chooses action $w(a)$, but R cannot always trust H to be rational.

2.1 Hadfield-Menell et. al.'s Approach

Hadfield-Menell et al. (2016b) model the off-switch game as a cooperative game. The human follows a policy π^H which models how rational H is. It is a function mapping U_a to a number $p \in [0, 1]$ representing the probability that H lets R do a. They denote R's belief state as B^R, which in this case is a distribution for U_a. The expected value of B^R given action a means the value R expects from taking the action. The variance of B^R represents R's uncertainty about what utility the intended action will generate. Hadfield-Menell et al. analyses this model with respect to the expected value and variance of B^R, and different kinds of policies π^H. Based on these parameters, they investigate the incentive, Δ, to choose $w(a)$:

$$\Delta = \mathbb{E}[\pi^H(U_a)U_a] - \max\{\mathbb{E}[U_a],\ 0\} \tag{1}$$

Equation (1) represents the difference in expected value for the robot between asking H and not asking H. When $\Delta \geq 0$, R has incentive to choose $w(a)$. When $\Delta < 0$, R will take action a if $\mathbb{E}[U_a] > 0$ and s otherwise. Given that H is rational they prove that $\Delta \geq 0$ regardless of what R's belief state is. They also show that if U_a follows a Dirac distribution, i.e. R is certain about U_a, then Δ is positive if and only if H is *rational*. We define rationality as in Definition 2.

Definition 2 (Rational). *A human H with utility function u is* rational *if H always picks action $a_H \in \{s, \neg s\}$, such that $a_H = \operatorname{argmax}_a u(a)$.*

Definition 2 means that R believe H to be rational if $\pi^H = 1$ if $U_a \geq 0$ and $\pi^H = 0$ otherwise, we denote this policy as π^H_r. The more interesting case when H is *irrational* is also analysed. The robot's belief distribution over U_a is assumed to be normally distributed. The irrationality of H is modelled with the sigmoid function (Eq. (2)), where β is a parameter controlling the degree of irrationality of H.

$$\pi^H(U_a; \beta) = \frac{1}{1 + e^{-\frac{U_a}{\beta}}}. \tag{2}$$

The degree of rationality of H increases as the parameter β tends towards zero in the policy function defined above. When β tends to infinity, $\pi^H(U_a; \beta)$ tends towards a completely random policy which takes action s and $\neg s$ with equal probability. We have that $\lim_{\beta \to 0} \pi^H(U_a; \beta) = \pi^H_r$ and $\lim_{\beta \to \infty} \pi^H(U_a; \beta) = \frac{1}{2}$.

The result from the analysis by Hadfield-Menell et al. (2016b) was that in order for R to be useful, there has to be a fine balance between the robot's uncertainty about H's utility function and H's rationality. If the robot is too certain about what H wants, and it knows H to be irrational, then it will have less incentive to let H switch it off. If, on the other hand, R is too uncertain, then R will have a strong incentive to choose action $w(a)$, but it will be too inefficient to be useful for H.

3 Game-Theoretic Approach

The analysis of the off-switch game by Hadfield-Menell et al. is not fully game theoretic since H is not strictly rational in their setup, which contradicts the axiom of rationality in game theory. Our goal in this section is to construct a game-theoretic model that is suitable for modelling the off-switch game. The idea is to represent an irrational human H as a rational agent H_r where the utility function of H_r is a modified version of H's utility function.

3.1 Modelling Irrationality

Since game theory is based on interaction between rational agents, we propose an alternative representation of the human in this subsection. We show that every irrational

human H can be represented by a rational agent maximising a different utility function. This allows us to use game-theoretic tools when analysing the off-switch game.

In general H is stochastic. R will believe H to be rational with some probability p.

Definition 3 (p-rational). *A human H with utility function u is* p-rational *if H picks action $a_H \in \{s, \neg s\}$ such that $a_H = \operatorname{argmax}_a u(a)$ with probability $p \in [0, 1]$.*

Since any type of irrationality boils down to a probability of making a suboptimal choice, p-rationality is a general model of irrationality.

Proposition 4 (Representation of irrationality). *Let H be a* p-rational *agent with utility function u, choosing between two actions s and $\neg s$. Then H can be represented as a rational agent H_r maximising utility function u with probability p and utility function $-u$ with probability $1 - p$.*

Proof. According to Definition 3, H is p-rational if it picks $a_H = \operatorname{argmax}_a u(a)$ with probability p and sub-optimal action $a'_H \neq a_H$ with probability $1 - p$. Since H only has two actions available, we have that $a'_H = \operatorname{argmin}_a u(a)$. This is therefore equivalent to maximising a utility function u with probability p and utility function $-u$ with probability $1 - p$. □

Proposition 4 states that a p-rational human can be modelled as a rational agent with random function. The proposition is a special case of a Harsanyi transformation (Rasmusen 1994).

3.2 Game-Theoretic Model

In this subsection we use the Harsanyi transformation, and Proposition 4 to model a p-rational human H as a rational agent H_r. This will allow us to model the off-switch game as an extensive form game between the rational players R and H_r. Nature N makes some moves that model R's uncertainty and these moves result in four leaves, each of which is a 3×2 strategic game between R and H_r.

We model the off-switch game by using the Harsanyi transformation a second time to let Nature choose the type of the rational human by choosing the utility function of the rational human after it has chosen the value of U_a. The resulting tree is represented in Fig. 2.

Definition 5 (The off-switch game). *A formal definition of our setup of the off-switch game is as follows.*

Players: *A robot R, a human H and Nature N. H's type is unknown to R, that is R does not observe Nature's moves.*

Order of Play:

1. *Nature chooses utility U_a that R generates from taking action a.*
2. *Nature decides the utility function of H, u^{H_r}, i.e. whether H is rational.*
3. *R chooses between actions in action set $\{a, w(a), s\}$.*
4. *If R chose $w(a)$ then H chooses between actions in action set $\{s, \neg s\}$.*

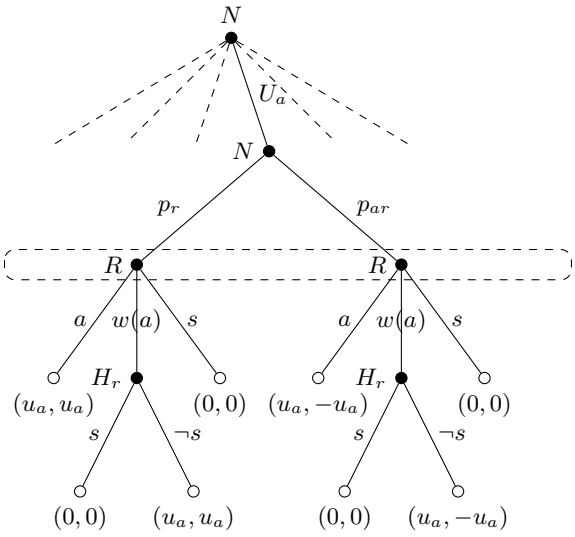

Fig. 2: Tree representation of the Off-Switch game after the second Harsanyi transformation. The nodes inside the dashed rectangle belong to the same information set. p_r is the probability that H_r has the same utility function as R and p_{ar} is the probability that H_r has the additive inverse of $R's$ utility function.

Note that unlike Hadfield-Menell et al. we view the off-switch game as a non-cooperative game. We find this reasonable since conflict arises when the robot and the human have different ideas about what is good for H. If the robot believes H is too irrational to be able to decide what is good for the human, R will not want to let H decide what to do even if R's purpose is to maximize H's payoff.

3.3 Aggregation

In this subsection we aggregate the branches in Fig. 2. This results in the game tree in Fig. 3, with four possible scenarios that can result from N's choices. The aggregation is possible since strategic play is never affected by positive linear transformations of the payoffs, hence the outcome of the games will only depend on the sign of U_a. We can therefore simplify the model by aggregating all branches of N's choices of U_a which has the same sign. This means that N has only two choices when deciding the utility U_a, that is if $U_a \geq 0$ or $U_a < 0$. The trivial case where $U_a = 0$, both R and H_r are indifferent about their actions and we will without loss of generality regard this case as U_a being positive.

We define R's subjective belief about N's aggregated choices as *primary statistics*. By primary statistics we mean parameters that are necessary to analyse our model. We also define the expected value of U_a as a primary statistics. This leaves us with a total of five primary statistics that are sufficient and necessary to model the off-switch game.

Primary Statistics 6. *Let the primary statistics $p_u^+ = P(U_a \geq 0)$ be the probability that U_a is positive. The event $U_a < 0$ is the complement of the event $U_a \geq 0$ and therefore we define $p_u^- = 1 - p_u^+$ as an auxiliary statistic.*

R's belief about H's rationality will depend on U_a. If $U_a \geq 0$ then the robot will believe H to be rational with probability p_r^+ and anti-rational with probability p_{ar}^+. If,

on the other hand, $U_a < 0$, the robot will believe H to be rational with probability p_r^- and anti-rational with probability p_{ar}^-. We define the following probabilities as primary statistics.

Primary Statistics 7. *Let the primary statistics $p_r^+ = P(H \text{ is rational} \mid U_a \geq 0)$ and $p_r^- = P(H \text{ is rational} \mid U_a < 0)$ be the probabilities that H is rational given that U_a is positive and negative respectively. The auxiliary statistics $p_{ar}^+ = 1 - p_r^+$ and $p_{ar}^- = 1 - p_r^-$ are the complementary probabilities that H is anti-rational.*

Primary Statistics 8. *Let the primary statistics $e_u^+ = \mathbb{E}[U_a \mid U_a \geq 0]$ and $e_u^- = \mathbb{E}[U_a \mid U_a < 0]$ be the expected value of U_a given that U_a is positive and negative respectively.*

From the perspective of R, N's choices can result in essentially four different subgames, denoted G_r^+, G_{ar}^+, G_r^- and G_{ar}^- illustrated in Fig. 3. In Fig. 4 we represent these subgames as 3×2 strategic games between two rational players; R, the robot, and H_r, a rational human.

The utility function, and hence the payoffs of R in the four games in Fig. 4 are determined by U_a. The utility function of H_r, on the other hand, is determined by the combination of U_a and the rationality type of H. H_r is always a rational agent in these games, i.e. H_r always maximises his expected payoff. H_r and R can be considered to have the same payoffs in each outcome if H_r has utility function u^{H_r} and the games G_r^+ and G_r^- associated with these scenarios are therefore no-conflict games. If on the other hand H_r has utility function $-u^{H_r}$ the payoff of H_r is the additive inverse of $R's$ payoff in each outcome. Therefore the games G_{ar}^+ and G_{ar}^- can be modeled as zero-sum games.

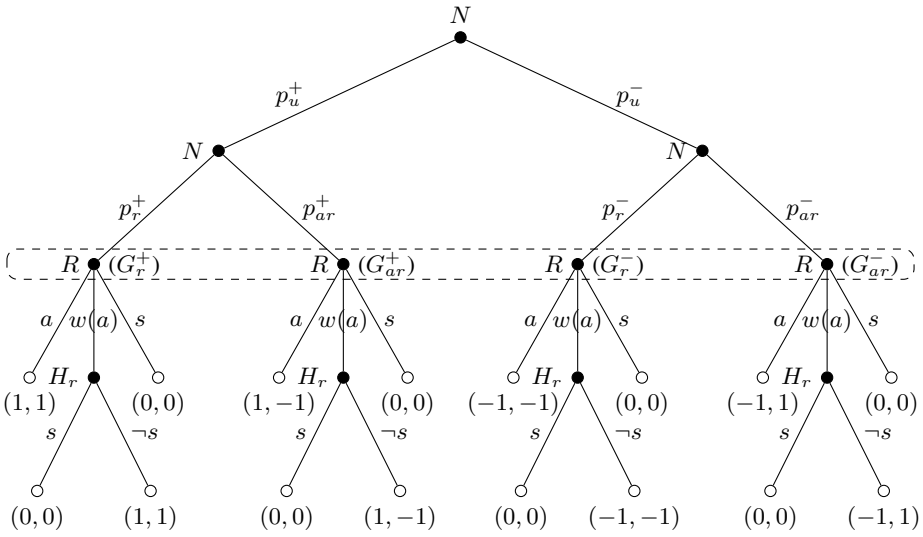

Fig. 3: Tree representation of the Off-Switch game after Harsanyi transformation. The nodes inside the dashed rectangle belong to the same information set. The subtrees denoted G_r^+, G_{ar}^+, G_r^-, G_{ar}^- are presented in strategic form in Fig. 4.

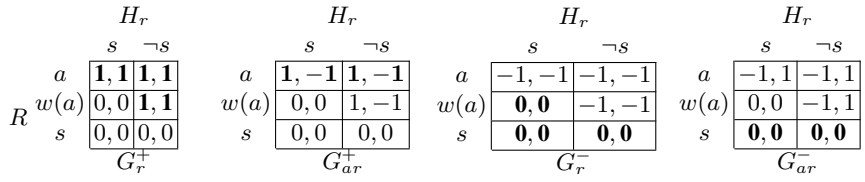

Fig. 4: The structure of the strategic games $G_r^+, G_{ar}^+, G_r^-, G_{ar}^-$. The outcomes with bold payoffs are Nash equilibria in each game.

For example in the scenario where $U_a < 0$ and the human is rational, the human will always choose s. Therefore in G_r^- the payoffs of H_r is aligned with the payoffs of R. Thus, if R chooses to take action $w(a)$, H_r prefers to take action s. In contrast, in the scenario where $U_a < 0$ and the human is irrational, H will choose the action $\neg s$. In other words, the payoffs of R and H_r are not aligned in the subgame G_{ar}^-.

3.4 Best Action

After having constructed the the game matrix, it is natural to now look at the expected value of each action using these matrices. The expected value for each action can be calculated as the expectation over all the possible subgames $G_r^+, G_{ar}^+, G_r^-, G_{ar}^-$ the robot can find himself in.

Theorem 9 (Main theorem). *The expected value of the actions for the robot are*

$$\begin{aligned} \mathbb{E}[U|s] &= 0 \\ \mathbb{E}[U|a] &= p_u^+ e_u^+ + p_u^- e_u^- \\ \mathbb{E}[U|w(a)] &= p_u^+ p_r^+ e_u^+ + p_r^- p_u^- e_u^- \end{aligned} \tag{3}$$

Proof. We compute the expected utility of the actions:

$$\begin{aligned} \mathbb{E}[U|s] &= 0 + 0 + 0 + 0 = 0 \\ \mathbb{E}[U|a] &= P(U_a \geq 0)\mathbb{E}[|U_a| \,|U_a \geq 0] + P(U_a < 0)\mathbb{E}[-|U_a| \,|U_a < 0] \\ &= p_u^+ e_u^+ + p_u^- e_u^- \\ \mathbb{E}[U|w(a)] &= P(r, U_a \geq 0)\mathbb{E}[U_a \,|U_a \geq 0] + P(\neg r, U_a < 0)\mathbb{E}[U_a \,|U_a < 0] \\ &= p_u^+ p_r^+ e_u^+ + p_{ar}^-(1 - p_u^+)e_u^- \\ &= p_u^+ p_r^+ e_u^+ + p_{ar}^- p_u^- e_u^- \end{aligned}$$

□

The expected value for taking the action s is 0, as we would expect from the definition of the off-switch game. The expected value for taking action a only uses information about the distribution of U_a, and like action s does not have any reliance on the human's rationality. It is a direct application of the law of total expectation. The expected value of action $w(a)$ is the difference between a positive term $p_u^+ p_r^+ e_u^+$ and a negative term $p_r^- p_u^- e_u^-$, both resulting from the human taking action a. The positive

term is the gain when U_a is positive and the human takes the action. The negative term is the loss when U_a is negative, and the human takes the action anyway (due to irrationality). The expected utility of $w(a)$ thus depends on the likelihood of U_a being positive (p_u^+) and the likelihood of human rationality (p_r^+), as well as the expected gains (e_u^+) and losses (e_u^-) in the respective cases.

Writing in this form allows us to come up with a useful corollary.

Corollary 10 (Compare a and $w(a)$). *Action a is preferred to $w(a)$ if and only if*

$$-p_u^+ p_r^+ e_u^+ + p_u^- p_r^- e_u^- > 0 \tag{4}$$

and the robot is indifferent if (4) *is equal to 0.*

Proof.

$$\begin{aligned}
(4) &= -p_u^+ p_r^+ e_u^+ + p_u^- p_r^- e_u^- \\
&= -p_u^+ p_r^+ e_u^+ + p_r^- e_u^- (1 - p_u^+) \\
&= -p_u^+ p_r^+ e_u^+ + p_u^+ e_u^+ + p_r^- e_u^- - p_u^+ p_r^- e_u^- \\
&= -p_u^+ p_r^+ e_u^+ - e_u^- + p_u^+ e_u^- + p_r^- e_u^- - p_u^+ p_r^- e_u^- + p_u^+ e_u^+ + e_u^- - p_u^+ e_u^- \\
&= -p_u^+ p_r^+ e_u^+ - (1 - p_r^-)(1 - p_u^+) e_u^- + (p_u^+ e_u^+ + (1 - p_u^+) e_u^-) \\
&= \mathbb{E}[U|a] - \mathbb{E}[U|w(a)]
\end{aligned}$$

If $\mathbb{E}[U|a] - \mathbb{E}[U|w(a)] > 0$ then $\mathbb{E}[U|a] > \mathbb{E}[U|w(a)]$ which occurs if and only if action a is preferred over $w(a)$. When (4) equals 0 then $\mathbb{E}[U|a] = \mathbb{E}[U|w(a)]$, hence the agent is indifferent. □

This provides us with a convenient way of testing for any distribution of U_a and r, and whether action a is preferred over $w(a)$.

4 Conclusion

In this paper, we have given a complete characterisation of how the robot will act in off-switch game situations for arbitrary belief and irrationality distributions. As established in our main Theorem 9, the choice depends only on 5 statistics. This result is much more general and arguably more useful than the one provided in the original paper (Hadfield-Menell et al. 2016b), as normal and soft-max assumptions are typically not realistic assumptions.

Off-switch game models an important dynamic in what we call the uncertainty approach to making safe agents, where the agent can choose to defer a decision to a human supervisor. Understanding this dynamic may prove important to constructing safe artificial intelligence.

Acknowledgements

This work grew out of a MIRIx workshop, with Owen Cameron, John Aslanides, Huon Puertas also attending. Thanks to Amy Zhang for proof reading multiple drafts. This work was in part supported by ARC grant DP150104590.

References

Allais, Maurice (1953). "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine". In: *Econometrica* 21.4, pp. 503–546. DOI: 10.2307/1907921.

Armstrong, Stuart (2015). "Motivated Value Selection for Artificial Agents". In: *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 12–20.

— (2010). *Utility Indifference*. Tech. rep. Oxford University, pp. 1–5.

Armstrong, Stuart and Jan Leike (2016). "Towards Interactive Inverse Reinforcement Learning". In: *NIPS Workshop*.

Dewey, Daniel (2011). "Learning what to Value". In: *Artificial General Intelligence*. Vol. 6830, pp. 309–314. ISBN: 978-3-642-22886-5. DOI: 10.1007/978-3-642-22887-2. arXiv: 1402.5379.

Everitt, Tom et al. (2016). "Self-modificication of Policy and Utility Function in Rational Agents". In: *Artificial General Intelligence*. Springer, pp. 1–11.

Hadfield-Menell, Dylan et al. (2016a). "Cooperative Inverse Reinforcement Learning". In: arXiv: 1606.03137.

— (2016b). "The Off-Switch Game". In: 2008, pp. 1–11. arXiv: 1611.08219.

Martin, Jarryd, Tom Everitt, and Marcus Hutter (2016). "Death and Suicide in Universal Artificial Intelligence". In: *Artificial General Intelligence*. Springer, pp. 23–32. arXiv: 1606.00652.

Omohundro, Stephen M (2008). "The Basic AI Drives". In: *Artificial General Intelligence*. Ed. by P. Wang, B. Goertzel, and S. Franklin. Vol. 171. IOS Press, pp. 483–493.

Orseau, Laurent and Stuart Armstrong (2016). "Safely interruptible agents". In: *32nd Conference on Uncertainty in Artificial Intelligence.*

Rasmusen, Eric (1994). *Games and Information*. 2nd ed. Blackwell.

Soares, Nate and Benja Fallenstein (2014). *Aligning Superintelligence with Human Interests: A Technical Research Agenda*. Tech. rep. Machine Intelligence Research Institute (MIRI), pp. 1–14.

Soares, Nate et al. (2015). "Corrigibility". In: *AAAI Workshop on AI and Ethics*, pp. 74–82.

Von Neumann, John and Oskar Morgenstern (1947). *Theory of Games and Economic Behavior*. Ed. by Lambert Schneider and Odette Deuber. Princeton Classic Editions. Princeton University Press. ISBN: 0691003629. DOI: 10.1177/1468795X06065810.

Wiener, Norbert (1960). "Some Moral and Technical Consequences of Automation". In: *Science* 131.3410, pp. 1355–1358. ISSN: 0036-8075. DOI: 10.1126/science.132.3429.741.