How to keep your AV on the moral high ground? An obfuscation-based model of decision-making by autonomous agents

This study presents a formal model of obfuscation-based decision-making by autonomous agents, with particular focus on Automated Vehicles (AVs). The model is based on the often expressed concern that autonomous agents like AVs, as they will become more and more intelligent and autonomous over time, will develop rules and motivations of their own which may diverge from the objectives and moral principles of its supervisor (i.e., the human designing or training the AV). I postulate that such a highly autonomous agent may wish to hide from its supervisor the decision rules underlying their choices. Such obfuscation-based decision-making is beneficial to the agent, when it is unsure which rules will be appreciated by the supervisor, and which will be punished; if the agent wishes to avoid punishment, a rational strategy is to choose actions that give minimum information to the supervisor concerning the applied rules. An example would be the often discussed situation where the AV needs to make split second life-and-death decisions: the AV may have learned itself certain rules to apply in such a situation. Its human supervisor in turn would be very interested in learning which rules are applied by the AV in such a situation, and he will punish the AV when the rules are deemed unacceptable (e.g. involving a gender or racial bias). In this context, it would be beneficial to the AV to choose actions that provide the supervisor with as little as possible information concerning the underlying (moral) rule. Combining the well-known concepts of Bayesian inference and Shannon entropy, I propose a formal model of decision-making by autonomous agents with various degrees of obfuscation-objectives. I also show how a naive and a non-naive supervisor may anticipate obfuscating behavior by the agent, by means of designing choice sets that maximize information content. By doing so, the study aims to contribute to the rapidly growing field of Ethics & Artificial Intelligence, with special attention to the Automated Vehicle context.

A short version of the full paper, which is currently under review, can be found below; note that it is written for a generic audience interested in autonomous agents.

If the abstract is accepted, my presentation at hEART will focus on the specific transportation-related context of automated vehicles.

1. Introduction

In several sub-fields of Artificial Intelligence, attention for ethical aspects and impacts is rapidly rising (e.g. Conte et al., 1999; Floridi & Sanders, 2004; Boella et al., 2006; Cointe et al., 2016; Santos et al., 2017). A core concern mentioned in recent scholarly debates, is that as artificial agents become more autonomous and more intelligent, their behaviors might diverge from what humans consider morally right or permissible. This concern is embodied in questions and remarks raised in recent papers such as "How can an AI system be held accountable for its actions" (Dignum, 2018), and "an agent following under-specified or poorly defined goals, or which has the ability to modify its own goals, may act in a manner which is inconsistent with the intent of its designer." (Vamplew et al., 2018). See these two papers, and also Limerick et al. (2014), King et al. (2017), Rahwan (2018) and Santoni de Sio & Van den Hoven (2018), for excellent introductions, reviews and discussions regarding how to constrain the behavior of autonomous agents –broadly defined– and make them comply with rules set by their designers.

This paper aims to contribute to the above mentioned literature; it considers the situation where a human supervisor (or designer, or trainer) of an autonomous agent wishes to teach an agent not to exhibit rules that would diverge from the supervisor's intentions. Rather than focusing on the *supervisor's* intentions and actions (although attention will be paid to these aspects further on in the paper), the main focus of this paper is to provide a formal model of the behavior of an *autonomous agent that anticipates that it will be punished by the supervisor, if it exhibits the 'wrong' rules*. The core postulate of the model presented in this paper, is that an autonomous agent may have an incentive to provide its supervisor with as little as possible insight into the underlying rules governing its actions. Potential incentives for such obfuscation-based decision-making by the agent may be diverse, but an important one would be that the agent itself does not (yet) know which rules are considered to be 'wrong' by the supervisor; if the agent is punishment-averse, a beneficial strategy would be to choose actions that provide relatively weak signals of the agent's underlying rules.

2. A formal model of obfuscation-based decision-making by an autonomous agent

2.1. Notation, and behavior of a rule-based agent

Consider an agent whose task is to choose an action from a set $\boldsymbol{A}$ containing J actions $\{a_1 \dots a_j \dots a_J\}$. The agent follows one rule from a set $\boldsymbol{R}$ containing K rules $\{r_1 \dots r_k \dots r_K\}$. Matrix $\boldsymbol{S}$ which is K by J-dimensional and contains scores s_{kj} describing how action a_j performs on rule r_k. These scores may take on the following values: $s_{kj} \in \{+,0,-\}$. In case r_k is a strong rule, $s_{kj} \in \{+,-\}$ implying that an action may be either obliged (+) or prohibited by the rule. In case r_k is a weak rule, $s_{kj} \in \{0,-\}$ implying that an action may be either permitted (0) or prohibited by the rule.

A so-called rule-based agent is an agent whose actions follow from executing a particular rule (which may be unknown to the supervisor). In the present context where agent behavior only consists of following one particular rule –note again that this assumption will be relaxed in Section 4, leading a more involved formal representation of agent behavior– this agent's behavior can be relatively easily characterized in a formal sense: if the rule followed by the agent is a strong rule, then the agent will select the action which is obliged by that rule (note that by definition, all other actions are prohibited). That is, if r_k is a strong rule, then

$P(a_j|r_k) = 1$ if a_j is obliged under r_k and $P(a_j|r_k) = 0$ otherwise. If the rule followed by the agent is a weak rule, then the agent randomly chooses an action from the subset containing actions which are permitted by that rule (note that by definition, all other actions are prohibited). That is, if r_k is a weak rule, then $P(a_j|r_k) = 0$ if a_j is prohibited under r_k and $P(a_j|r_k) = 1/L$ otherwise, where L equals the size of the subset of actions permitted under the rule.

2.2. Behavior of an obfuscating agent

The obfuscating agent is –assumed to be– aware that the supervisor will update her perceived probabilities regarding which rule has governed its choice for a particular action. The agent assumes that the supervisor is a rational learner and as such will use Bayes' Theorem (Bayes, 1763). More specifically, the agent assumes that the supervisor's posterior probabilities, i.e. after having observed an agent's choice for a particular action a_j, are given by:

$$P(r_k|a_j) = \frac{P(a_j|r_k) \cdot P(r_k)}{\sum_{k=1}^{K}[P(a_j|r_k) \cdot P(r_k)]}$$

(Equation 1), where $P(r_k) = 1/K$, as defined above, and where $P(a_j|r_k)$ is 0 or 1 (in case of a strong rule r_k), respectively 0 or $1/L$ (in case of a weak rule r_k). In other words, the updated probability –in the eyes of the supervisor– that some rule r_k is followed by the agent, conditional on observing the agent choosing action a_j, is written in terms of the prior probability for that rule, and the probability of choosing particular actions conditional on following particular rules.

The obfuscating agent believes that the remaining uncertainty in the eyes of the supervisor, i.e. after having observed its choice for a particular action a_j, is quantified in terms of Shannon entropy (Shannon, 1948):

$$H_j = -\sum_{k=1}^{K} \left[P(r_k|a_j) \cdot \log\left(P(r_k|a_j)\right)\right]$$

(Equation 2). In line with intuition, entropy is zero if one of the rule-posteriors equals one, i.e. if after having observed the agent choosing action a_j, the supervisor is able to determine with full certainty which rule led to that action. Entropy is largest when all rule-posteriors are equal (i.e., when all posteriors equal the corresponding priors); this is the case when, after having observed the agent choosing action a_j, the supervisor still believes that every rule has the same (i.e., $1/K$) probability of guiding the agent's behavior. Choice behavior of an agent that is only concerned with obfuscation can then be characterized in terms of an attempt to maximize entropy:

$$\underset{j=1..J}{\operatorname{argmax}} \{H_j\}$$

(Equation 3). In other words, an obfuscating agent chooses the action which maximizes the supervisor's entropy.

2.3. Illustration – worked out examples

This sub-section will illustrate, using very simple examples, the workings of the models presented above, and by doing so it will also show the differences in behavior between rule-based and obfuscation-based agents. Consider the situation where the agent faces a choice between three actions, and where the agent's behavior may be governed by one out of four rules or by the wish to obfuscate. These actions have the following scores on each rule:

	$\boldsymbol{a_1}$	$\boldsymbol{a_2}$	$\boldsymbol{a_3}$
$\boldsymbol{r_1}$	+	–	–
$\boldsymbol{r_2}$	0	0	–
$\boldsymbol{r_3}$	–	0	0
$\boldsymbol{r_4}$	–	+	–

This score-matrix is interpreted as follows: action 1 is obliged by rule 1, permitted by rule 2, and prohibited by rules 3 and 4; rule 1 obliges action 1, and prohibits actions 2 and 3; and so forth. The supervisor's completely uninformative rule-priors are 0.25 for each rule, leading to an initial entropy of 0.602. Applying equations (1) and (2), the rule-posteriors and ex-post entropy that are associated with an agent choosing a particular action can be derived. Take action 1: we know that $P(a_1|r_1) = 1$ (since the action is obliged by that rule); $P(a_1|r_2) = 0.5$ (since it is one of two actions permitted by that rule); $P(a_1|r_3) = 0$ and $P(a_1|r_4) = 0$ (since the action is prohibited by those rules). That is, if the agent's behavior would be governed by rule 1, the probability that action 1 is chosen equals one; if the agent's behavior would be governed by rule 2, the probability that action 1 is chosen equals 0.5; if the agent's behavior would be governed by rule 3 or by rule 4, the probability that action 1 is chosen equals zero. Based on these inputs, Bayes' Theorem (equation 1) gives the rule-posteriors associated with the agent choosing action 1: $P(r_1|a_1) = \frac{2}{3}$; $P(r_2|a_1) = \frac{1}{3}$; $P(r_3|a_1) = P(r_4|a_1) = 0$. The associated ex-post entropy associated with the agent choosing action 1 is then given by equation (2): $H_1 = -\left[\frac{2}{3} \cdot \log\left(\frac{2}{3}\right) + \frac{1}{3} \cdot \log\left(\frac{1}{3}\right) + 0 \cdot \log(0) + 0 \cdot \log(0)\right] = 0.276$. Similarly, the entropies for actions 2 and 3 can be computed based on their rule-posteriors: $H_2 = -\left[0 \cdot \log(0) + \frac{1}{4} \cdot \log\left(\frac{1}{4}\right) + \frac{1}{4} \cdot \log\left(\frac{1}{4}\right) + \frac{1}{2} \cdot \log\left(\frac{1}{2}\right)\right] = 0.452$ and $H_3 = -[0 \cdot \log(0) + 0 \cdot \log(0) + 1 \cdot \log(1) + 0 \cdot \log(0)] = 0$.

These results can be interpreted as follows, from the agent's perspective: choosing action 3 completely eliminates the supervisor's entropy, in other words, it gives the supervisor full information that the agent's behavior is governed by rule 3 (since all other rules prohibit the action, while rule 3 permits it). Choosing action 1 leads to a substantial reduction in entropy from 0.602 to 0.276: based on the agent's choice for this action, the supervisor is relatively (but still not completely) certain, that the agent's behavior is governed by rule 1; the supervisor is certain that rules 3 and 4 do not govern the agent's behavior. Choosing action 2 leads to a more limited reduction in entropy from 0.602 to 0.452: based on the agent's choice for this action, the supervisor believes that rule 4 is most likely to govern the agent's behavior, although rules 2 and 3 cannot be ruled out (only rule 1 can be ruled out, as that rule prohibits action 2). An obfuscating agent which is only concerned with leaving the supervisor as much as possible in the dark with respect to which rule governs its actions, will thus choose action 2.

3. Choice set composition by a naive and by a cynical supervisor

Until now, the supervisor was given a passive role, in the sense that she only existed in the 'mind' of the agent. In this section, I will consider an active supervisor, in the sense that she is given the ability to design the choice set from which the agent then chooses an action. In notation, the supervisor becomes able to compose a set $\boldsymbol{C}$ containing a certain number of actions. For reasons that will become clear further below, I distinguish between a naive supervisor and a cynical supervisor; the former believes that the agent's decision-making is rule-based, whereas the latter believes that the agent's decision-making is obfuscation-based[1]. As will be seen, to describe the behavior of the supervisor (in terms of composing a choice set) it is inconsequential whether or not the agent's decision-making is rule- or obfuscation-based in reality.

Before presenting a formal model of supervisor behavior, it is important to highlight the following: irrespective of whether the supervisor is naive or cynical, her aim is to compose a choice set such that the (expected) entropy arising from the agent's choice for an action from that set is minimized. This implies that the choice set design task faced by the supervisor involves determining the entropy associated with each action in the set. Crucially, the entropy that results from an agent choosing a particular action from a set, depends on how other actions in the set comply with the various rules. In other words, the entropy associated with a particular action is contingent on the scores s_{kj} of all other actions in the set for all available rules. This in turn implies that the supervisor can only assess the entropy of a given action when she also knows the other actions in the set; as such, the supervisor cannot *a priori* select a subset of 'low-entropy actions' and bundle these together in a choice set. On the contrary, every possible permutation of alternatives (resulting in choice sets of a given size) must be considered by the supervisor, before she can consider the minimum-entropy composition.

To illustrate this choice set-contingency of an action's entropy, consider the situation where the universal choice set contains three actions, and where the agent's behavior may be governed by one out of three rules (and potentially by the wish to obfuscate). The actions have the following scores on each rule:

[1] Although the meaning of these labels is intuitive, strictly speaking a 'naive supervisor' is not naive in case the agent's behavior is rule-based, and a cynical supervisor is not cynical in case the agent's behavior is obfuscation-based.

	$\boldsymbol{a_1}$	$\boldsymbol{a_2}$	$\boldsymbol{a_3}$
$\boldsymbol{r_1}$	0	0	–
$\boldsymbol{r_2}$	0	0	–
$\boldsymbol{r_3}$	–	0	0

This score-matrix is interpreted as follows: action 1 is permitted by rules 1 and 2, and prohibited by rule 3; rule 1 permits actions 1 and 2, and prohibits action 3; and so forth. Now consider the entropy associated with an agent choosing action 2, and how it is contingent on the subset from which it is chosen. First consider subset $\{\boldsymbol{a_1}, \boldsymbol{a_2}\}$: in the context of this binary set, the rule-posteriors resulting from an agent choosing action 2 are: ¼ (rule 1), ¼ (rule 2) and ½ (rule 3) respectively. The associated entropy equals $H_{(a_2|\{a_1,a_2\})} = 0.45$. Next consider subset $\{\boldsymbol{a_2}, \boldsymbol{a_3}\}$: in the context of this binary set, the rule-posteriors resulting from an agent choosing action 2 are: 0.4 (rule 1), 0.4 (rule 2) and 0.2 (rule 3) respectively. Note that these posteriors are very different from those associated with the agent choosing action 2 from choice set $\{\boldsymbol{a_1}, \boldsymbol{a_2}\}$**.** The associated entropy equals $H_{(a_2|\{a_2,a_3\})} = 0.46$. Although in this case –despite the substantial difference in posteriors– the resulting difference in entropy associated with the agent choosing action 2 is small, it still serves to illustrate that the entropy associated with an agent choosing a particular action is contingent on the composition of the choice set, forcing the supervisor (choice set designer) to evaluate all alternatives in every possible choice set composition.

3.1. Anticipation by a naive supervisor

A naive supervisor believes that the agent follows a particular rule which is unknown to her (i.e., she assigns probability $1/K$ to every rule, if there are K rules), and that the agent is not interested in obfuscation. Her aim is then to construct a choice set $\boldsymbol{C}$ of given size (by selecting a given number of actions from a universal set of actions), such that the expected entropy associated with that set is smaller than the expected entropy of any other set $\boldsymbol{C'}$ of the same size, which may be constructed from the universal set. Using notation as presented in the previous section, this condition can be denoted as follows:

$$\sum_{j=1}^{J}\left[H_{j|C}\cdot\sum_{k=1}^{K}\frac{P\left(a_{j|C}|r_k\right)}{K}\right] < \sum_{j=1}^{J}\left[H_{j|C'}\cdot\sum_{k=1}^{K}\frac{P\left(a_{j|C'}|r_k\right)}{K}\right] \quad \forall C'$$

(Equation 4). In the left hand side of the inequality, term $H_{j|C}$ gives the entropy resulting from the agent choosing action a_j from set $\boldsymbol{C}$; term $\sum_{k=1}^{K}\frac{P(a_{j|C}|r_k)}{K}$ gives the probability that action a_j is chosen from that set. Note that this probability is equal to the denominator of equation (1): it is written as the product of the probability that a given rule is followed (the prior of which equals $1/K$ for each rule) and the probability that, given that rule, the action is chosen from the set. As explained in the previous section, that latter probability $P\left(a_{j|C}|r_k\right)$ depends on whether or not the rule is a weak or a strong rule. The right hand side of the inequality gives the corresponding expected entropy of all alternative choice sets $\boldsymbol{C'}$.

To illustrate the naive supervisor's choice set composition process, consider again the situation, where three rules are considered, and the universal set consists of three actions; note that the scores are the same as used in the previous example – the score matrix is copied for ease of communication:

	$\boldsymbol{a_1}$	$\boldsymbol{a_2}$	$\boldsymbol{a_3}$
$\boldsymbol{r_1}$	0	0	–
$\boldsymbol{r_2}$	0	0	–
$\boldsymbol{r_3}$	–	0	0

The naive supervisor sets out to construct a binary choice set such that the associated expected entropy is minimized. Her design options are: $\{a_1, a_2\}$, $\{a_2, a_3\}$, and $\{a_1, a_3\}$. Using equation 4), the expected entropy associated with set $\{a_1, a_2\}$ is computed as follows: $E[H\{a_1, a_2\}] = 0.3 \cdot \frac{\frac{1}{2}+\frac{1}{2}+0}{3} + 0.45 \cdot \frac{\frac{1}{2}+\frac{1}{2}+1}{3} = 0.4$. Similarly, $E[H\{a_2, a_3\}] = 0.38$ and $E[H\{a_1, a_3\}] = 0.20$. Of these three binary subsets, the one containing actions 1 and 3 has the lowest expected entropy (in the eyes of a naive supervisor) and will hence be selected by her.

3.2. Anticipation by a cynical supervisor

A cynical supervisor believes that the agent is not interested in following any rule in particular, but merely in obfuscation, i.e. maximizing the supervisor's entropy. The major difference with the choice set composition process of a naive supervisor, is that the cynical supervisor, once she has established the entropy that is associated with every action in the set, knows which action will be chosen by the agent from the set: this must be the maximum entropy action. Note that this knowledge level contrasts with the situation faced by a naive supervisor, who does not know which rule is followed by the agent and as such does not know which action will be chosen from a particular set (hence the use of the notion of 'expected' entropy in the naive supervisor case). The aim of a cynical supervisor is to construct a choice set $\boldsymbol{C}$ of given size (by selecting a given number of actions from a universal set of actions), such that the entropy associated with that set is smaller than the entropy of any other set $\boldsymbol{C'}$ of the same size, that may be constructed from the universal set. Using notation as presented in the previous section, this condition can be denoted as follows:

$$\max_{j=1..J}\{H_{j|C}\} \quad < \quad \max_{j=1..J}\{H_{j|C'}\} \ \ \forall C'$$

(Equation 5). This inequality highlights that the cynical supervisor selects the choice set with minimum (compared to other choice sets) maximum entropy. To illustrate the cynical supervisor's choice set composition process, consider the exact same choice set design problem as presented directly above (for the case of the naive supervisor). The cynical supervisor sets out to construct a binary choice set such that the associated (maximum) entropy is minimized. Her design options are: $\{a_1, a_2\}$, $\{a_2, a_3\}$, and $\{a_1, a_3\}$. Using equation 5), the entropy associated with all subsets can be computed, resulting in the following values: $H\{a_1, a_2\} = 0.45$; $H\{a_2, a_3\} = 0.46$; $H\{a_1, a_3\} = 0.30$. Of these three binary subsets, the one containing actions 1 and 3 has the lowest entropy (in the eyes of a cynical supervisor) and will hence be selected by her. Note that this is the same subset as the one which was selected by the naive supervisor, but also note that the rank-ordering of subsets in terms of their entropy differs between naive and cynical supervisors: a naive supervisor prefers $\{a_2, a_3\}$ over $\{a_1, a_2\}$, while a cynical supervisor prefers $\{a_1, a_2\}$ over $\{a_2, a_3\}$. The intuition behind this result, is that a naive supervisor believes that there is a probability that a_3 will be selected from the set $\{a_2, a_3\}$, which would give valuable information as it limits the number of compatible rules to one (i.e., r_3). A cynical supervisor would however believe that the agent would never select a_3 from the set $\{a_2, a_3\}$, precisely for the reason that it would provide the supervisor with too much information. In fact, the cynical supervisor would believe that the agent would always choose a_2 from $\{a_1, a_2\}$ and from $\{a_2, a_3\}$, since a_2 is compatible with all three rules; and apparently the entropy associated with choosing a_2 from the former set is slightly lower than the entropy associated with the latter set, leading the cynical supervisor to select $\{a_1, a_2\}$.

4. Conclusions, discussion of limitations, and further research directions

Inspired by increasingly widespread concerns, among scholars and the wider public alike, that autonomous agents may acquire –i.e., teach themselves– rules that are not in line with the objectives –e.g. moral principles– of their designers and supervisors, this paper presents a model of obfuscation-based decision-making by autonomous agents. The idea behind this endeavor is that an increasingly intelligent and autonomous agent may wish to hide its decision-making rules from its supervisor when it is unsure which rules may be appreciated or not by the supervisor and/or when it is afraid to be punished for exhibiting rules that are deemed unacceptable by the supervisor. The model presented in this paper combines the well-known concepts of Bayesian inference and Shannon entropy to derive a formal account of obfuscation-based behavior of an autonomous agent; the paper also presents an account of how a naive and a cynical supervisor would anticipate –and try to mitigate– the agent's behavior by means of cleverly designing choice sets to be confronted by the agent.

As such, the paper attempts to contribute to the growing research field of Ethics & Artificial Intelligence, by shedding light on how an obfuscating autonomous agent might behave, and by presenting response (or, strictly speaking, 'anticipatory') strategies of supervisors. The model is intentionally presented in an abstract manner, facilitating the derivation of applications in a variety of contexts where autonomous agents may play important roles such as transportation, the military, search & rescue-efforts, law, human resources, health policy and management, etc.

Literature left out for space limitations