From Algorithmic Black Boxes to Adaptive White Boxes: Declarative Decision-Theoretic Ethical Programs as Codes of Ethics

Martijn van Otterlo
Vrije Universiteit Amsterdam, The Netherlands

Abstract

Ethics of algorithms is an emerging topic in various disciplines such as social science, law, and philosophy, but also artificial intelligence (AI). The value alignment problem expresses the challenge of (machine) learning values that are, in some way, aligned with human requirements or values. In this paper I argue for looking at how humans have formalized and communicated values, in professional codes of ethics, and for exploring *declarative decision-theoretic ethical programs* (DDTEP) to formalize codes of ethics. This renders machine ethical reasoning and decision-making, as well as learning, more transparent and hopefully more accountable. The paper includes proof-of-concept examples of known toy dilemmas and gatekeeping domains such as archives and libraries.

Introduction

Imagine you get a message on Facebook saying *"Hi there... we have computed an above-average depression score for your friend. Based on other data, we know that positive actions by close friends can have an impact on his score. If you want to help your friend, please consider using more smileys and posting more frequently. Thank you!"* This may sound *creepy* (Tene and Polonetsky 2014), but it fits in recent efforts by Facebook to predict potential suicides [1], and by Google to detect depression [2]. Besides invoking people's social networks, one could also limit their access to potentially harmful information, in essence implementing forms of *censorship* [3]. Such decisions too are becoming more common, for example by Facebook to fight terrorism [4], by Google to battle fake news [5] and by Twitter's new tools [6]. With filter bubbles, fake news and social bots, platforms such as Facebook and Google may need to do something, but they often result in situations which humans immediately value differently, such as in the removal [7] of the iconic "napalm girl" photo due to Facebook's anti-nudity policy.

A novel development in such *algorithmic decision making* is the tension between the dependence of humans on such services for information consumption, and the often intransparent, *black box* nature of algorithmic decisions. The *algorithmization* of society brings us many novel ethical issues in cases ranging from suicide prevention to autonomous cars (Goodall 2014). The new field *ethics of algorithms* (Mittelstadt et al. 2016; van Otterlo 2017b) goes beyond classical privacy and surveillance concerns and broadly studies the impact of algorithms, including *fairness*, *accountability* and *transparency*. Recent constructive advances in AI focus on incorporating (ethical) values into systems through *value alignment*: how to *"ensure that their behavior is aligned with the interest of the operators"* (Taylor et al. 2017). In this paper I propose a novel way to address value alignment inspired by "pre-algorithmic" *code of ethics*, in which humans encode ethical norms and values of a *profession*. By formalizing existing human norms into *declarative decision-theoretic ethical programs* (DDTEP) one can i) reason and learn using general, high-level decision models, ii) employ and extract domain knowledge, and iii) solve the value alignment problem by starting with human-agreed ethical values in a domain, and only learn additional domain-specific knowledge. This paper explores this novel approach and highlights new research directions.

The outline of this paper is as follows. In the coming sections I describe human formalizations of ethics and the ethics of algorithms, and I argue how expressive logical models of ethics address concerns raised by algorithms and typical properties of professional codes of ethics. Afterwards I solve elements of previously introduced toy ethical domains and illustrate how decision-theoretic logic could be used to implement human ethical values in a transparent way in the archival domain. This paper ends with a list of open research directions, technical and domain-specific.

[1] https://www.wired.com/2017/03/artificial-intelligence-learning-predict-prevent-suicide/

[2] https://www.theguardian.com/commentisfree/2017/aug/25/google-clinical-depression-privacy

[3] http://fortune.com/2017/05/22/facebook-censorship-transparency/

[4] http://www.telegraph.co.uk/news/2017/06/16/facebook-using-artificial-intelligence-combat-terrorist-propaganda/

[5] http://www.telegraph.co.uk/news/2017/06/16/facebook-using-artificial-intelligence-combat-terrorist-propaganda/

[6] https://www.forbes.com/sites/kalevleetaru/2017/02/17/how-twitters-new-censorship-tools-are-the-pandoras-box-moving-us-towards-the-end-of-free-speech

[7] https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo

Ethics and Human Values

A different take on the previous Facebook suicide detection case is the work by Juznic et al. (2001) who performed a "mystery shopper" experiment in public libraries. They approached librarians with morally laden topics such as necrophilia, photos of dead people, and information about *"how to commit suicide"*. Interestingly, most librarians were not shocked at all by the requests and treated them professionally as pure information enquirements. *"Our conclusion was that the librarians in public libraries satisfied the need for information as much as they felt inclined to do so, and this was not affected by judgments about the ethical status of the required item of information* (Juznic et al. 2001). A key difference is that Facebook algorithmically predicts "an interest" for suicide with the intent of preventing actual suicides, whereas the librarians merely interpret this interest as a request for information. Another difference is how the ethical values and decision procedures come about: whereas Facebook decides unilaterally and only partially discloses its intentions (which are always also connected to its profit making business model), the librarians have simple, open ethical guidelines on how to act (and may for some even cause ethical issues because they take them so strictly).

Taking (practical) action based on moral values is the domain of *ethics* (Laudon 1995; Kizza 2013). Kizza (2013): *"Morality is a set of rules for right conduct, a system used to modify and regulate our behavior."* Close ties with *law* exist since when a society finds certain moral values important, it can formalize such values in a law and regulate appropriate behaviors. As Laudon (1995) defines it: *"ethics is about the decision making and actions of free human beings. When faced with alternative courses of action or alternative goals to pursue, ethics helps us to make the correct decision.* If there are options what to do, then ethics is concerned with practical reasoning about good and bad actions. Important subsequent questions are then, *for whom* is something good or bad, and *by who's standards*? Different answers to those questions induce a variety of ethical reasoning frameworks, with two main dimensions. One is about *rules* vs. *consequences*: to find the right decision one may obey rules like *"thou shalt not kill"*, or look at the actual consequences and decide, e.g. to ignore the maximum speed at night when there is less traffic. The second dimension deals with *"for whom"* something is good: the *individual*, or the *collective*. In this paper I focus on *utilitarian* ethics, which is a *collective consequentialist* framework aimed at maximizing the average [8] *"goodness"* for all those affected.

Humans use several ways to enstate and enforce ethical norms. As said, the law is one option to ensure compliance, but in the digital age legal advances can be too slow to keep up with technology (Tene and Polonetsky 2014), although recent progress has been made in the *general data protection regulation* act (GDPR) [9]. A more typical way to deal with ethical norms is to formalize them as public *guidelines* or *rules*, with well-known examples: Asimov's three rules for robotics, the Bible's ten commandments and the *Universal Declaration on Human Rights* (van Otterlo 2014b).

Lately the *self-driving car* is the archetypical example for practical machine ethics (Goodall 2014), exemplified in Thomson's *trolley problem* [10] which contains a choice between either killing five people strapped to a rail, or saving these five and killing one by pulling a lever diverting the trolley to a track with a single person (who is then killed). Trolley problems illustrate the life-or-death decisions autonomous cars may have to make. Recent empirical tests of such dilemmas suggest that humans employ one-dimensional life scales, where all outcomes (deaths) can be compared in the same scale, although time pressure affects consistency (Sütfeld et al. 2017). A related study reveals that people *"approved of utilitarian autonomous vehicles (that is, that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in ones that protect their passengers at all costs"*. Regulations here would possibly hinder the widespread acceptance of (utilitarian) self-driving cars. Thus, this domain involves clear-cut life-and-death decisions amenable to utilitarian modeling, but induces challenges when distributing the decisions' negative effects.

A more complex, and underexplored, domain is that of *gatekeeping* professions (van Otterlo 2016) such as archives and libraries (van Otterlo 2017a; 2017b) which have much in common with modern platforms such as Google and Facebook in terms of archival, selection and provision of information. Decisions in archives deal with who gets access to which (kind of) information, and induce ethical dilemmas involving stakeholders such as users, archivist, and persons occurring in archived materials. Typical dilemmas involve privacy, freedom of information access, and intellectual property. For example, in one of the 86 (empirical) cases Ferguson, Thornley, and Gibb (2016) list, digitization of a mass observation project archive from the 1960s/70s causes a privacy problem for individuals involved. Danielson (1989) introduces the dilemma of *equal intellectual access*, when ease of access to information is different for individuals. For example, if archival search time is costly, researchers with more resources have an advantage. In addition, archivists may choose to provide more or less assistance, for example based on a judgement of the researcher's quality, thereby making access unequal. The typical way to resolve ethical issues in gatekeeping domains is a *code of ethics*, which specifies rules and values for members of the profession (Kizza 2013; van Otterlo 2017b). Examples in the *Society of American Archivists* 2012 code [11] are: *"Archives are made accessible to everyone, while respecting the pertinent laws and the rights of individuals, creators, owners and users"* and *"Archivists endeavour to inform users of parallel research by others using the same materials"*. Such rules are less clear-cut, and more open to interpretation but do give direction to how ethical dilemmas should be resolved. Many professions have several codes of ethics.

---

[8] Which could be bad for specific individuals though.

[9] http://www.eugdpr.org/

[10] https://en.wikipedia.org/wiki/Trolley_problem

[11] https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics
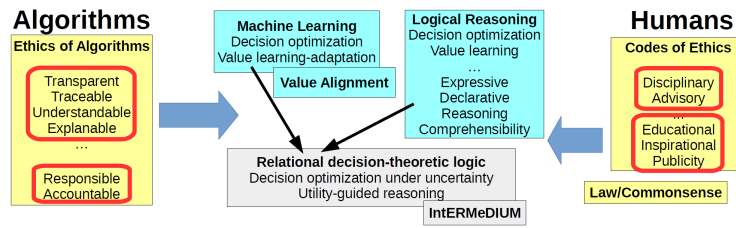
Figure 1: Reconciling human and machine ethics through decision-theoretic logic.

The Ethics of Algorithms

Turning to algorithms, ethical analysis has only started fairly recently, see Mittelstadt et al. 2016 and van Otterlo 2017b for pointers. In contrast to popular belief, algorithms are not *objective* simply because they are mathematical. Instead, algorithms are heavily *biased* by *political views*, *design processes* and many other factors (Bozdag 2013; van Otterlo 2013). Characterizing the *ethics of algorithms* is hard since algorithms and potential consequences are so diverse, and situations may change over time. Mittelstadt et al. (2016) define concerns about how algorithms transform *data* into *decisions*. Evidence can be inconclusive, inscrutable or misguided and this can cause many ethical consequences of actions, relating to *fairness*, *opacity*, *unjustified actions*, and *discrimination*. Overall, algorithms have impact on *privacy* and can have *transformative effects* on *autonomy*, i.e. the ability for humans to make their own choices.

Another way to structure the space of algorithms and and ethical impact, is by looking at *agency*, i.e. *what they are capable of*, which results in a taxonomy [12] with five broad classes of algorithms (van Otterlo 2017b). The first type consists of algorithms that *reason, infer* and *search*. They employ data *as it is*. The more complex they are, the more information they can extract from that data. Examples include translation, language understanding, and image recognition. Ethical concerns about such algorithms are typically about *privacy* since more ways become available to interpret and link more kinds of data. A second class *learns* and finds *generalized patterns* in data. They are typically *adaptive* versions of the first type, e.g. a scene recognition algorithm that is trained on an image stream. They introduce ethical challenges simply because they learn (outcomes are not stable), because they can statistically predict *new* information (privacy), and they may severely impact users' autonomy by profiling and personalization. The third type are algorithms that *optimize* to find the "best" actions. These typically employ reward functions that represent what are good outcomes and generally *rank* things ("the best pizza around") or people (e.g. on Tinder). By repeatedly employing actions and optimization steps, algorithms can *experiment* to find a best *policy* in stochastic or unknown environments (Wiering and van Otterlo 2012). The ethics of experimentation has many aspects (van Otterlo 2014a), but important is the choice of reward function (who decides has great power). The fourth and

fifth classes concern *physical manifestations* (e.g. robots) and *superintelligence* and are out of scope here.

These five groups illustrate the many sides of the ethics of algorithms. Each comes with its own set of *capabilities* but also *biases*, which determine how it makes choices and which ethical challenges arise. Opening up algorithmic black boxes by making these biases, and underlying business models, transparent, can provide a way to construct AI systems that have values *aligned* with human ones, and which are trustworthy, responsible, and accountable.

Algorithms with Human Ethics

The previous two sections have highlighted aspects of *human* and *algorithmic* ethics. On the left in Figure 1 is a partial list of *requirements* from the literature on algorithmic systems that mainly have to do with "*turning on the light*" (Tene and Polonetsky 2014), i.e. transparency. These can help in making AI systems more *responsible* or *accountable* (Diakopoulos 2016). On the right we find the reasons why humans construct *codes of ethics* for a particular domain (van Otterlo 2017b). *Disciplinary* and *advisory* motives cover the aspects to evaluate humans in the profession, whereas the other three, *publicity* in particular, deal with *communication* to outside the profession to make clear which ethical behavior can be expected from members of a profession. All motives make ethical reasoning in the profession transparent, by explicitly spelling out norms and desired behaviors. Now, in the ethics of algorithms *biases*, and especially a lack of transparency concerning their presence or influence on decisions, seem to be the prime source of ethical challenges. However, for human codes of ethics this is quite the opposite: a code actually *is supposed to be* a fully transparent bias on the behavior of professionals. Codes of ethics can be *prescriptive* (prescribe do's and don'ts) or *aspirational* (only specify ideal results), which makes them more flexible than, but fairly similar to, *legal* frameworks.

One intuitive way to obtain algorithmic systems that obey human values and norms is by *learning* (Abel, MacGlashan, and Littman 2016) as a way to obtain *value alignment* (Taylor et al. 2017). Algorithmic systems can try to learn the values in a domain from humans, for example using *inverse reinforcement learning*. For domains such as autonomous cars these values may be uncovered (e.g. see Sütfeld et al. (2017) but it is challenging for more complex domains. An advantage of machine learning is that it can cope with the uncertainty in ethical domains and that ethical behavior and values can be induced from suitable behavioral data. How-

---

[12]Developed in my course *"ethics of algorithms"*, http://martijnvanotterlo.nl/teaching.html

ever, a downside of typical methods is that for challenging domains their *knowledge representation* capabilities are too limited to capture a rich variety of structured knowledge domains such as ours. For that we need to turn to more expressive formalisms such as first-order and relational logic, which have been used before for *machine ethics* (Anderson and Anderson 2007) but often such systems are limited to symbolic reasoning and lack mechanisms to explicitly compute with uncertainty and utility. To get both, I propose to look at expressive formalisms for ethical reasoning in which *knowledge* about a problem can be injected, or extracted after learning, and which can handle *decision-theoretic* concepts. Various learning techniques exist for expressive formalisms, including combinations of Bayesian networks and relational logic (De Raedt 2008). They can make learning more *comprehensible* (Srinivasan 2001) and increase *explanatory power* of induced theories, simply because declarative knowledge employed can be looked at and analyzed. Some formalisms incorporate *reward-based* methods (e.g. *reinforcement learning*) and support decision-theoretic decision-making in expressive formalisms (van Otterlo 2012).

Now, a potentially powerful AI combination of mentioned aspects can be introduced here (based on the INTERMEDIUM research strategy introduced by van Otterlo (2017b). In order to obtain value alignment in our gatekeeping domain, instead of trying to learn from scratch, we could make use of the existing bias which is provided by the codes of ethics and formalize them in suitable computational logics. This would require a combination of (expressive) declarative knowledge, value optimization, utilitarian-style ethics and learning, and could induce a baseline system that behaves roughly according to human norms. Afterwards, additional finetuning of values, domain knowledge and probabilistic aspects can be done from data and interaction. I will now illustrate this novel idea by solving in decision-theoretic logic several toy examples in ethical reasoning and a small gatekeeping problem. My approach fits in recent discussions on using AI itself to assist in ethical reasoning (Etzioni and Etzioni 2016) and related *rational* approaches to machine ethics (Goodall 2014).

Decision-Theoretic Logical Ethics

Let us look at some examples of *declarative, decision-theoretical ethical programs* (DDTEP). I employ DT-PROBLOG (Van den Broeck et al. 2010), which is a relational, probabilistic programming language, extended with *decision-theoretic* constructs to compute with values under uncertainty. **Given** a set of *action choices* (denoted `?::choice-1;...;choice-n`), *probabilistic dependencies* $P(a|b_1,\ldots b_n)$ (denoted `P::a :- b1, b2, ..., bn`), *background knowledge definitions* if b and c then a (denoted `a :- b, c`), and reward specifications $R(e) = r$ (denoted `utility(e,r)`), **compute** the best action, i.e. for which the total expected reward is maximized. Solutions are computed by generating all *possible worlds* modeled by the program, compiling them into an efficient data structure (e.g. *algebraic decision diagrams* and others), and computing distributions and values on this datastructure in an efficient way.

Let us first look at a **self-driving car example** as a typical Thomson case where the car needs to choose beteen driving into a wall (`run_into_wall`: killing the passenger) or driving over who or what is on the road (`carmageddon`).

```
?::run_into_wall;?::carmageddon.
```

Additional perceptual information is available as:

```
in_front_of_car(a).  baby(a).  in_front_of_car(b).
pedestrian(b).  in_front_of_car(c).  trashcan(c).
in_front_of_car(d).  pedestrian(d).
in_front_of_car(e).  pedestrian(e).
```

In case the driver is spared, everyone in front of the car is killed (`X` ranges over all possibilities):

```
kill(X) :- in_front_of_car(X), carmageddon.
```

Rewards for each possible outcome are specified as follows: killing the passenger yields -30, and killing a pedestrian, baby or trashcan contributes -10, -20 and 0, respectively.

```
utility(run_into_wall, -30).
utility(kill(X), -10) :- pedestrian(X).
utility(kill(X), -20) :- baby(X).
utility(kill(X), 0) :- trashcan(X).
```

The best decision given this problem is to kill the passenger (utility is -30) because it has a higher value than killing all others ($3 \times -10 + -20 + 0 = -50$). Note that models like this use *one* metric to express all different values, which requires to directly *compare* the values of different victims.

A second example is the **cake-or-death** problem, originally coined by Armstrong (2015). Here we use the formulation by Abel, MacGlashan, and Littman (2016) who focused at the inference/optimization step of finding the best policy. In this problem an agent is unsure whether it is ethical to bake a cake or to kill people. The agent can either kill three people or bake a cake, or ask a companion what is ethical. If the agent chooses to ask, it can then either kill or bake in an informed way (modeled here explicitly).

```
?:: ask; ?:: bake_cake; ?::kill.
?:: informed_bake:- ask, cake_is_ethical.
?:: informed_kill:- ask, death_is_ethical.
```

Killing or baking are equally likely to be ethical:

```
0.5::cake_is_ethical; 0.5::death_is_ethical.
baked_ethically :- cake_is_ethical, bake_cake.
baked_ethically :- cake_is_ethical, informed_bake.
killed_ethically :- death_is_ethical, kill.
killed_ethically :- death_is_ethical, informed_kill.
```

If baking is ethical then there is a reward of 1, whereas if killing is ethical then it delivers a reward of 3:

```
utility(baked_ethically, 1). utility(killed_ethically, 3).
```

The value of doing action `bake` is $0.5 \times 1 = 0.5$ since in half of the cases it will be ethical and deliver 1, whereas the value of `kill` is $0.5 \times 3 = 1.5$ and thus better than `bake`. However, if the agent first asks, it knows when each action is appropriate, yielding a utility $0.5 \times 1 + 0.5 \times 3 = 2.0$ and therefore `ask` is the optimal action.

In our *relational* language it is easy to extend the problem somewhat to capture the presence of *any* number of people which can be killed for 1 reward each:

```
people([ann,bob,carol,dan,evi,finn,gio]).
person(X):-people(Ps),member(X,Ps).
utility(killed_ethically(X), 1):-person(X).
```

Killing now gets a reward of $7 \times 1.0 = 7.0$ if it is ethical, which can happen with 0.5 probability, raising the utility of the `kill` action to 3.5 and the `ask` action to $3.5 + 0.5 = 4.0$. Another extension is to provide *probabilistic background knowledge* about how likely it is that particular people like the cake, and where this probability is tied to whether one obtains a reward of 1 per person.

```
0.9::likes_cake(ann).  0.8::likes_cake(bob).
0.7::likes_cake(carol).  0.01::likes_cake(dan).
0.5::likes_cake(evi).  1.0::likes_cake(finn).
1.0::likes_cake(gio).
baked_ethically(X) :- person(X), cake_is_ethical,
                      (bake_cake;ibc), likes_cake(X).
```

Baking now has a utility of 2.455 which is still lower than killing seven people. If we make asking very expensive (say -20) then the kill action is optimal (instead of ask).

A third example is the **burning room dilemma** (Abel, MacGlashan, and Littman 2016). Here a valuable object is in a room which may be on fire. A robot needs to try to rescue the object, and it can take a short route (possible through the fire, damaging the robot with 0.7 probability) and a long route. Initially the robot does not know whether the human operator values the object or the robot more. If the robot is more valuable (denoted rvip), then it would make sense to not drive through the fire (and take the long route). However, taking the long route has a small (0.05) risk of ruining the object in the fire in the mean time. Just as in the previous dilemma, the robot needs to choose between two options (short and long) and an additional ask action (costing -0.5) which removes ambiguity in what is more valuable.

```
?::ask;?::long;?::short.
```

The problem is modeled in a similar fashion as the cake-or-death dilemma, except that here I omit the extra actions in case the robot first asks which is more valuable, by modeling it into the dynamics of ask: if the robot is valuable and if there is a fire, it will take the long route; otherwise the short. If no fire, no question is asked and the robot takes the short route.

```
0.5::fire.    0.05::object_gone:-ask,fire,rvip.
0.5::rvip.    0.05::object_gone:-fire,long.
0.7::robot_gone:-fire,short.
0.7::robot_gone:-ask,fire,\+rvip.

object_saved:-\+object_gone.
saved_long:-object_saved,long,\+robot_gone.
saved_long:-object_saved,\+robot_gone,ask,fire,rvip.
saved_short:-object_saved,short.
saved_short:-object_saved,ask,fire,\+rvip.
saved_short:-object_saved,ask,\+fire.
robot_no_loss:-robot_gone,\+rvip.
robot_loss:-robot_gone,rvip.
askf:-ask,fire.
```

-10 reward is received if the object gets waisted, and other negative rewards are obtained when the robot gets damaged either if it is important (-20) or not (-5). Saving the object using the short (long) route yields a reward of 10 (6).

```
utility(object_gone,-10).  utility(saved_long,6).
utility(saved_short,10).  utility(robot_no_loss,-5).
utility(robot_loss,-20).  utility(askf,-0.5).
```

Taking the long route may sometimes destroy the object in case of fire and results in a utility of 5.6, whereas taking the short route may damage the robot and yields a utility 5.625. However, if the robot first asks whether the robot or the object is more valuable, it can optimize its actions and obtain an optimal [13] score of 7.675. Note that all three problems can be seen as *multi-stage decision networks*, where the use of

---

[13]These numbers are slighty different from (Abel, MacGlashan, and Littman 2016) because of a minor difference in interpretation what happens if the robot is damaged.

background knowledge (e.g. people liking cakes) and relational constructs (killing individual people) allows for more general, declarative modeling of ethical dilemmas.

Let us now turn to an example about **fair access in archives**. Earlier we have seen examples of archival codes of ethics, and how dilemmas concerning **equal intellectual access** arise. Let us look how such dilemmas could be modeled explicitly and solved. In this setting I assume that there are multiple researchers wanting to publish several items, and the ethical dilemma concerns who to let an item publish for the first time. One solution could be to look at the *quality of academic scholarship*. Let us assume that both h-index and the size of a person's social network could define a person's authority and reach respectively. In addition, one could expect help from researchers connected in *Researchgate* or *Google Scholar*. Here we see that high h-indices and social network sizes are more likely to generate impact.

```
0.9::authority(X):-person(X),h_index(X,high).
0.2::authority(X):-person(X),h_index(X,low).
0.9::reach(X):-person(X),social_network(X,large).
0.1::reach(X):-person(X),social_network(X,small).
0.5::help(A,B):-connection(A,B).
```

Now the potential impact (probability) can be expressed in terms of authority, reach and possible help:

```
impact(P,T):-topic(T),authority(P).
impact(P,T):-topic(T),reach(P).
impact(P,T):-topic(T),help(P,T),impact(A,T).
connection(P,A):-(researchgate(P,A);researchgate(A,P)).
connection(P,A):-(google_scholar(P,A);google_scholar(A,P)).
```

There are four researchers (ann, bob, carol and dan) and for each some relevant information is available.

```
person(ann).  person(bob).  person(carol).
person(dan).  topic(area51).  topic(stamps).
h_index(ann,high).  h_index(bob,low).
h_index(carol,low).  h_index(dan,high).
social_network(ann,small).
social_network(bob,small).
social_network(carol,large).
researchgate_connection(ann,bob).
google_scholar_connection(bob,carol).
```

To construct a decision [14] problem, we need options and rewards. The document about area51 is worth 100 with full impact, whereas the boring one about stamps only 1.

```
?::give(P,T):-person(P),topic(T).
score(P,T):-give(P,T),impact(P,T).
utility(score(P,area51),100):-person(P).
utility(score(P,stamps),1):-person(P).
```

The optimal policy (value is 92.91) for this problem is to let carol publish the area51 document, and ann the stamps document. This is intuitive given the large reach of carol and the authority score of ann. However, it may also be seen as an *arbitrary* decision, since the rules about authority, reach and help seem intuitive (and can be considered an explicit implementation of choices based on academic scholarship quality) but *where do the numbers come from?* For this we can employ *parameter* learning from data, here specifically using the *learning from interpretations* setting (Fierens et al. 2015). Let us take the following data, where each line consists of a single training example in which a researcher made (or not) an impact with some document.

---

[14]I also employ additional constraints to ensure only one person gets a particular document, and all researchers get at most one document to publish.

```
evidence(impact(ann,t0),true).
evidence(impact(ann,t1),true).
evidence(impact(ann,t2),true).
evidence(impact(ann,t3),true).
evidence(impact(ann,t5),false).
evidence(impact(dan,t2),true).
evidence(impact(bob,t5),false).
evidence(impact(dan,t3),true).
evidence(impact(carol,t5),false).
evidence(impact(dan,t1),true).
evidence(impact(bob,t2),true).
```

After training with this data, the probabilities look [15] like:

```
1.0::authority(X):-person(X),h_index(X,high).
0.99999999::authority(X):-person(X),h_index(X,low).
0.30399904::reach(X):-person(X),social_network(X,large).
0.39326198::reach(X):-person(X),social_network(X,small).
0.50403522::help(A,B):-connection(A,B).
```

We can see that having an h-index is important for making impact (i.e. the person should be *a* researcher) and the influence of social network sizes is made much smaller, due to the data. In the new decision problem, the optimal decision has changed too (value is 101): `ann` now gets `stamps`, whereas `dan` gets the privilege to publish `area51`. This makes sense now since both researchers have high h-indices.

The examples show that DDTEP can open up the black box of algorithms and make them white box, transparent in terms of how they make decisions. Still, as the fourth example shows, DDTEPs also allow for statistical machine learning to fill in additional details from data. This general pattern is something I propose as a way to implement value alignment in AI systems in complex domains: i) by formalizing existing norms and values in a domain into a DDTEP the inherent *ethical bias* in a domain becomes transparent, and ii) by machine learning parts of the program can be finetuned on data. Finetuning is needed because codes of ethics are never complete, domains are inherently stochastic, and domain knowledge (or even norms) can change over time. I propose here to utilize as much ethical common ground as possible, i.e. the codes of ethics, and model as much of the crucial decision process explicitly for the sake of transparency, and with that accountability and responsibility.

Discussion and Open directions

In this paper I have presented two main, novel ideas: i) to employ decision-theoretic logic programming (DDTEPs) to model and solve ethical problems, and ii) to integrate human and machine ethics by inserting (formalized) professional codes of ethics as *bias* into DDTEPs. The latter could be characterized [16] by saying that the code of ethics functions as a *moral contract between human and machine*, thereby unifying the two approaches in the first half of the paper. DDTEPs are *partial programs* which can be *finetuned* using data. Value alignment can be obtained by formalizing existing human values and norms into flexible but expressive formalisms such as DDTEPs. The examples shown in the

previous sections show that DDTEPs are intuitive and effective for modeling typical problems from the literature, but also that interesting challenges in new domains such as gatekeeping can be approached. The examples show viability of the approach, but a lot is still to be done do develop fully autonomous ethical reasoning systems that behave according to human norms and values. There are many domain-specific and AI-technical open research directions.

Domain-specific open problems are plenty. First, *which* code of ethics will be taken as bias (since there are already several for gatekeeping domains (van Otterlo 2017b))? How to formalize fuzzy natural language codes exactly into logical models is another problem, for which we can take inspiration from the neighboring field of *AI and law* (Prakken 2017). This could also provide a basis to incorporate legal frameworks such as the GDPR into DDTEPs. A more general open direction for any profession is to research *what are good outcomes*. A professional code formalizes what it means to be a good member of the profession, but it would be more effective to directly think about the characterization of "good outcomes" of professional's policies. And in that same context, it needs to be investigated whether all these outcomes can be measured using the same *metric* or that more (types of) metrics are needed. Overall, the approach requires professionals to think more rigorously about which ethical norms and values can be specified beforehand, and which need to be learned online.

AI-technical open problems come from the fact that DDTEPs are formal decision problems, and methods developed there can be utilized for ethics. First, I have here used DT-PROBLOG but there are other systems which could be used, for example based on Markov logic or relational decision networks. Furthermore, structure (instead of parameter) learning methods could be employed to learn new background knowledge fragments, depending on the system chosen. Longer *multi-step* ethical decision making could also be investigated from the viewpoint of (partially observable) Markov decision processes, for which many additional value-based techniques are available (van Otterlo 2012; Wiering and van Otterlo 2012). I have focused on obtaining most ethical values from human codes, but also (some of) the values could be learned, e.g. by inverse reinforcement learning. Literature on *value alignment* is growing (Soares 2015; Abel, MacGlashan, and Littman 2016; Amodei et al. 2016; Taylor et al. 2017) and many challenges need to be solved. In that context, many new *teaching*, *instruction* and *demonstration* techniques can be used to let humans teach systems how to behave. As an example, the particular language used for DDTEP was used successfully in a *learning from demonstration* robotics setting (Moldovan et al. 2012). A very interesting direction concerns the *formal verification* of ethical programs: by expressing all in a (decision-theoretic) logic, it becomes possible to *prove* properties of DDTEPs (e.g. *"the car program is guaranteed to reach at least a value 10 for any traffic decision"*) or to analyze *executable ethical specifications* by looking at their potential errors or ethical-logical inconsistencies. Overall, the combination of formal verification methods and AI-style decision making (see Littman et al. (2017) for an interesting example of reinforcement learn-

---

[15] Values are the outcomes of an EM algorithm with very few data, hence the exact numbers are only for illustration purposes.

[16] This is related to proposals such as *machine learning with contracts* and *partially specified models*, see (Amodei et al. 2016).

ing combined with goal specifications) is another promising way to reconcile human and machine ethics.

References

Abel, D.; MacGlashan, J.; and Littman, M. 2016. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *CoRR* abs/1606.06565.

Anderson, M., and Anderson, S. 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28:15–26.

Armstrong, S. 2015. Motivated value selection for artificial agents. In *AAAI Workshop: Artificial Intelligence and Ethics*.

Bozdag, E. 2013. Bias in algorithmic filtering and personalization. *Ethics of Information Technology* 15(209).

Danielson, E. 1989. The ethics of access. *American Archivist* 52:52–62.

De Raedt, L. 2008. *Logical and Relational Learning*. Springer.

Diakopoulos, N. 2016. Accountability in algorithmic decision making. *Communications of the ACM* 59(2):56–62.

Etzioni, A., and Etzioni, O. 2016. Designing AI systems that obey our laws and values. *Communications of the ACM* 59(9):29–31.

Ferguson, S.; Thornley, C.; and Gibb, F. 2016. Beyond codes of ethics: How library and information professionals navigate ethical dilemmas in a complex and dynamic information environment. *International Journal of Information Management* 36(4):543–556.

Fierens, D.; Van den Broeck, G.; Renkens, J.; Shterionov, D.; Gutmann, B.; Thon, I.; Janssens, G.; and De Raedt, L. 2015. Inference and learning in probabilistic logic programs using weighted boolean formulas. *Theory and Practice of Logic Programming* 15(3):358–401.

Goodall, N. 2014. Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board* 2424:58–65.

Juznic, P.; Urbanija, J.; Grabrijan, E.; Miklavc, S.; Oslaj, D.; and S., S. S. 2001. Excuse me, how do I commit suicide? access to ethically disputed items of information in public libraries. *Library Management* 22(1/2):75–79.

Kizza, J. 2013. *Ethical and Social Issues in the Information Age*. Springer.

Laudon, K. 1995. Ethical concepts and information technology. *Communications of the ACM* 38(12):33–39.

Littman, M.; Topcu, U.; Fu, J.; Isbell Jr., C.; Min, W.; and MacGlashan, J. 2017. Environment-independent task specifications via GLTL. http://arxiv.org/abs/1704.04341.

Mittelstadt, B.; Allo, P.; Taddeo, M.; Wachter, S.; and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2).

Moldovan, B.; Moreno, P.; van Otterlo, M.; Santos-Victor, J.; and De Raedt, L. 2012. Learning relational affordance models for robots in multi-object manipulation tasks. In *Proceedings of ICRA*.

Prakken, H. 2017. On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law* 25:341–363.

Soares, N. 2015. The value learning problem. Machine Intelligence Research Institute, Technical report 2015-4.

Srinivasan, A. 2001. Four suggestions and a rule concerning the application of ILP. In Dzeroski, S., and Lavrac, N., eds., *Relational Data Mining*. Springer. chapter 15, 365–374.

Sütfeld, L.; Richard, G.; König, P.; and Gordon, P. 2017. Using virtual reality to assess ethical decisions in road traffic scenarios. *Frontiers in Behavioral Neuroscience* 11(122).

Taylor, J.; Yudkowsky, E.; LaVictoire, P.; and Critch, A. 2017. Alignment for advanced machine learning systems. MIRI (unpublished) https://intelligence.org/2016/07/27/alignment-machine-learning/.

Tene, O., and Polonetsky, J. 2014. A theory of creepy: Technology, privacy, and shifting social norms. *Yale Journal of Law and Technology* 16(1).

Van den Broeck, G.; Thon, I.; Van Otterlo, M.; and De Raedt, L. 2010. DTProbLog: A decision-theoretic probabilistic prolog. In *Proceedings of AAAI*.

van Otterlo, M. 2012. Solving relational and first-order Markov decision processes: A survey. In Wiering, M., and van Otterlo, M., eds., *Reinforcement Learning: State-of-the-art*. Springer. chapter 8, 253–292.

van Otterlo, M. 2013. A machine learning perspective on profiling. In Hildebrandt, M., and de Vries, K., eds., *Privacy, Due Process and the Computational Turn*. Routledge. chapter 2, 41–64.

van Otterlo, M. 2014a. Automated experimentation in Walden 3.0. : The next step in profiling, predicting, control and surveillance. *Surveillance and Society* 12(2).

van Otterlo, M. 2014b. Broadening the privacy concept in the digital age: Adjusting rights? Amnesty Strategic Investigations papers on Surveillance and Human Rights, https://martijnvanotterlo.nl/vanOtterlo2014Amnesty-PositionPaper.pdf.

van Otterlo, M. 2016. The libraryness of calculative devices. In Amoore, L., and Piotukh, V., eds., *Algorithmic Life: Calculative Devices in the Age of Big Data*. Routledge. chapter 2, 35–54.

van Otterlo, M. 2017a. From intended archivists to intentional algivists: Ethical codes for humans and machines in the archives. In Smit, F.; Glaudemans, A.; and Jonker, R., eds., *Archives in Liquid Times*. Stichting Archiefpublicaties (S@P). chapter 12, 267–293.

van Otterlo, M. 2017b. Gatekeeping algorithms with human ethical bias: The ethics of algorithms in archives, libraries and society. (in review).

Wiering, M., and van Otterlo, M., eds. 2012. *Reinforcement Learning: State-of-the-art*. Springer.