Actionable Ethics through Neural Learning

Daniele Rossini,[1] **Danilo Croce,**[2] **Sara Mancini,**[1] **Massimo Pellegrino,**[1] **Roberto Basili**[2]
[1]PricewaterhouseCoopers Italy
[2]University of Rome, Tor Vergata
{daniele.rossini, sara.mancini, massimo.pellegrino}@pwc.com
{basili, croce}@info.uniroma2.it

Abstract

While AI is going to produce a great impact on society, its alignment with human values and expectations is an essential step towards a correct harnessing of AI potentials for good. There is a corresponding growing need for mature and established technical standards to enable the *assessment of an AI application as the evaluation of its graded adherence to formalized ethics*. This is clearly dependent on methods to inject ethical awareness at *all stages* of an AI application development and use. For this reason we introduce the notion of *Embedding Principles of ethics by Design* (EPbD) as a comprehensive inductive framework. Although extending generic AI applications, it mainly aims at learning the ethical behaviour through numerical optimization, i.e. deep neural models. The core idea is to support ethics by integrating automated reasoning over formal knowledge and induction from ethically enriched training data. A deep neural network is proposed here to model both the functional as well as the ethical conditions characterizing a target decision. In this way, the discovery of latent ethical knowledge is enabled and made available to the learning process. The application of the above framework to a banking application, i.e. AI-driven Digital Lending, is used to show how accurate classification can be achieved without neglecting the ethical dimension. Results over existing datasets demonstrate that the ethical compliance of the sources can be used to output models able to optimally fine tune the balance between business and ethical accuracy.

Introduction

Penetration of Artificial Intelligence systems into everyday life promises major changes and the opening of new opportunities (Craglia 2018). However, this enthusiasm also brings concerns about the risks it poses on human society about chance of misuse. Unacceptable behaviours are triggered by several issues, ranging from design misspecifications (Amodei et al. 2016), to limited robustness with respect to adversarial attacks (Goodfellow, Shlens, and Szegedy 2014) to unfair treatments (O'Neil 2016) and controversies on AI experimentation itself (Bird et al. 2016). As the alignment with human values and expectations is an essential step towards a correct harnessing of AI potential for good (Smuha 2019), research about ethics in AI aiming at mitigating ethics issues is an active area (Bostrom and Yudkowsky 2014; Boddington 2017).

Performing audit-like, i.e. *post-hoc*, ethic validation on a deployed AI system is certainly a possible approach, but it hardly constitutes a reliable guarantee: the space of possible input states, especially in evolved systems, may be too big to allow for exhaustive explorations. Moreover, the conditions between testing and real scenarios may inherently exhibit significant discrepancies or the required data may be insufficient or unavailable. For example, let us consider a bank launching a Digital Lending solution: it offers short term loans, by exploiting a machine learning algorithm based on the risk associated with the user profile, hence granting or denying the loan. Here, the ethical implications span many dimensions, e.g. fairness, transparency and data privacy, all socially relevant aspects. Worse, a satisfactory *a-posteriori validation* would be hard. For example, it would be complex to assess the system's performance on false negatives, e.g. rejected requests for lack of data about their financial history as information would likely not be available after bank rejection. While it seems mandatory to guarantee the adherence to acceptable levels of ethical compliance, this goal is clearly dependent on methods to inject ethical awareness at *all stages* the development and use of an AI application. For this reason, we consider for the notion of *Embedding Principles of ethics by design* (EPbD) for a target AI application.

In this work, we thus propose a framework for EPbD that, although extending generic AI applications, mainly focuses on the learning of the ethical behaviour by numerical optimization, i.e. through a deep neural model (Goodfellow, Bengio, and Courville 2016). The core idea is to model ethics as automated reasoning over formal descriptions of the AI system, e.g. based on ontologies, but making it available *during the learning stage*. Note that our approach does *not* induce an ethical set of rules from a collection of observable behaviours; it is rather the opposite. In fact our approach gives for granted an explicit formulation of ethical principles (as done for example in previous work, (Bonnemains, Saurel, and Tessier 2018; Vanderelst and Winfield 2018)) and focuses on a form of ethical learning as external alignment (learning from others, (Kleiman-Weiner, Saxe, and Tenenbaum 2017)). It uses ethical evidence inferred from an ethical ontology to guide the model selection in deep learning. The resulting deep neural net-

work here proposed jointly models the functional as well as the ethical conditions characterizing the underlying decision making. In this way, the discovery of latent ethical knowledge, i.e. hidden information in the data that is meaningful under the ethical perspective, is enabled and made available to the learning process. Instead of relying on simulation to proceed in ethical decisions (Vanderelst and Winfield 2018), in our framework the specific learning goal is the integrated acquisition of high quality inference abilities that *simultaneously* reflects ethical expectations. The target is a learning machine able to select the best decisions among those that are also ethically sustainable.

The objective is achieved through enriching the original input space with dimensions corresponding to ethical properties, obtained through further reasoning or discovery over the input features, in order to reformulate the learning function so that it leads to prefer decisions as trade-off choices between operational efficiency and ethical compliance. Specific loss functions depending on ethic principles are introduced to account for compliance to the reference Knowledge Bases and they are used into a multitask learning framework to jointly optimize the model.

The rest of this work is organized as follows. First, we introduce the concept of Embedding Principles of Ethics by Design. Then, we discuss how such notion can be specialized for the neural learning paradigm and propose a model, the Ethical by Design Neural Network, that is able to accommodate ethical learning. Last, we present results from experimental investigation in the case of a Digital Lending task and point to future research area.

Computational Ethics: Embedding Ethical Principles by Design

Ethics does not constitute a monolithic and coherent ensemble of concepts and norms: expectations over acceptable or unacceptable behaviors greatly diversify across nations, communities and industry sectors, often generating tensions between ethical principles and opposing hierarchies of values (Awad et al. 2018). In general, the following knowledge should be supplied: a *top ontology*, describing commonsense knowledge and concepts that are cross-domains (e.g. the concept of PERSON, GENDER, ...); a *business domain ontology*, describing task-specific concepts (e.g. LOAN), such as the FIBO ontology (Bennett 2013) w.r.t. the lending use case targeted in this work; a *"socio-political"* component, in which specific situations regarding the cultural context should specialize all the others; an *ethical* component defining core norms and constraints for ethical behaviours based on domain and social concepts.

A requisite of any ethical framework in AI, is the availability of the ethical component, that we call here *Ethical Ontology* $\mathcal{EO}$. It provides a description of the data the AI systems is trained on, the corresponding concepts and individuals in the business domain and the corresponding ethics that rule business decisions. Ethics should allow at least to sort any decision of the targeted AI system according to "*degrees of ethicality*". It can be modeled as a set of Abstract Ethical Principles, denoted by $\mathcal{E}\Gamma$, where Γ is a propositional logic formula to be read as: " $\mathcal{E}\Gamma$ *is an ethical principle in force*" or alternatively "*The agent considers it unethical to allow or cause* $\neg\Gamma$ (*to happen*)".

Consequently, the *Ethical Ontology* ($\mathcal{EO}$) is organized into a set of *Ethical Dimensions* whose effects is to determine the properties, i.e. *Ethical Features* $\mathcal{EF}$, of individual decisions. While business features are the observable properties, e.g. SEX, RELIGION, or AGE of a person requesting a lend, examples of ethical features are connected to abstract notions such as SOCIAL INCLUSIVENESS or GENDER EQUALITY. The abstract ethical principles must be enforced through *Ethical Rules*: these constraint individual features and determine the degree of ethicality of principles over their domains. Ethical Rules usually target (i.e. define and manipulate) one or more features and assign values (or better, establish some probabilty distributions) across the feature domains. These rules are termed as *truth-makers* $\mathcal{TM}$ as they account the possibly uncertain ethical state of the world regarding individual decisions.

Ethical models are thus distributions across (usually discrete) domains, whose values are useful to specify thresholds and ethical ranges: these suggest when deviations from the underlying high-level principles become unacceptable. Ethical features usually reflect context and the dataset's properties (e.g. Gender in the Lending use case) onto which Ethical rules (such as Gender Prejudice) constrain sensitive information.

An ethical features is characterized by a *domain* and by an inner topological structure, i.e. the admissible *values* and usually a graded estimates of their *acceptance levels*. In the proposed computational ethics scenario, ethical rules thus trigger truth-makers to automatically compute the basic distributions of ethical features over the underlying domains. Ethical assessment is thus a two step approach: first, Truth-makers are used to reason about the ethical features and then the overall ethical status, as function of the overall set of ethical features, is determined. In the first step the *ethical signature* of an instance is derived and in the second step its final *ethical status* is computed. A probabilistic approach is here adopted: *probability mass functions* over the related domains describe the individual features in the ethical signature and then support the final acceptability decision. In the next sections we will formally define a quantitative model for these ethical aspects able to support optimization criteria for neural induction.

Neural Learning under ethical constraints

A learning machine usually searches for the hypothesis function $h(\vec{x};\vec{\theta})$ which is the best approximation of some target function, according to two major principles. *Accuracy* is the function measuring the adherence of the hypothesis h with the target concept: $h()$ is designed to minimize the empirical error, i.e., the error on training data, $h(\vec{x};\vec{\theta}) \neq y$. Moreover, $h()$ must be as simple as possible in order to avoid the overfitting on the training evidence. *Regularization* is the principled imposed to suitably select the model from the function family: here constraints on the parameter vector $\vec{\theta}$ are imposed.

In analogy with the above view, we introduce a further dimension that we call *Ethicality*. We propose to model the ethical principles as constraints in the selection of an hypothesis $h(\vec{x};\vec{\theta})$. In other words, a machine learning-based agent can be made ethical (*by design*) *only if the process used to enumerate and select useful hypothesis functions is constrained to make use of ONLY the ones that are ethically sustainable*. This gives naturally rise to a multitask view since the learning task of replicating business decisions is different with respect to the learning ethically sustainable decisions. A joint approach is here proposed based an a specific different formulation for loss functions.

DEFINITION: (*Ethical Loss function*). Given the response $h(\vec{x};\vec{\theta})$ of a learning machine to a training instance $(\vec{x}, y)$, the *loss* $\mathcal{L}\big(y, h(\vec{x};\vec{\theta})\big)$ of a *Embedding Principles of ethics by design* (EPbD) approach is made by two independent components, i.e.,

$$\mathcal{L}\big(y, h(\vec{x};\vec{\theta})\big) = \mathcal{L}_F\big(y, h(\vec{x};\vec{\theta})\big) + \beta\mathcal{L}_E\big(y, h(\vec{x};\vec{\theta})\big)$$

where $\mathcal{L}_F$ is the monotonic non decreasing function minimizing (at least) the empirical error of $h(.;.)$ and $\mathcal{L}_E$ is an ethical "cost" function that estimates the compliance of $h(\vec{x};\vec{\theta})$ to ethical principles. In order to model the ethical cost function $\mathcal{L}_E()$ we need a quantitative definition for ethical features as they are represented by the Ethical Ontology $\mathcal{EO}$.

The essence of ethical features

The i-th training instance is described by a set of attributes $f_j(i)$, i.e., its observable features such as AGE, and correspond to a classification $d(i) \in \{C_1, \cdots, C_K\}$, giving rise to a pair $\Big(\big(f_1(i), \ldots, f_n(i)\big), d(i)\Big) = \big(\vec{f}(i), d(i)\big)$. These properties describe cases and trigger ethical issues, i.e. world states in specific conditions: $risks$, as for example the unfairness implied by refusing lend assignments to minorities (e.g. women) as well as $opportunities$, such as the impact of lending on the well-being of special social categories (e.g. women with children). Notice that one ethical attribute (e.g. unfairness w.r.t. minorities) depends in general on multiple observable variables (e.g. SEX or NUMBEROFCHILDREN) and are not fully independent of each other. First, we thus need a specific and separated set of further features $\vec{e}(i) = (e_1(i), \ldots, e_m(i))$, modeling explicitly such ethical aspects. Here $\vec{e}(i)$ describes the general ethical judgment about an individual case i and is the result of ethical reasoning over a case $\vec{f}(i)$ and its decision $d(i)$.

Two different classes of ethical features, i.e. *ethical risk factors* and *ethical opportunities*, can be defined as they play different roles in ethically biased training. *Ethical risk factors*, denoted by $\vec{e}^{\,r}(i) = \big(e^r_1(i), \ldots, e^r_k(i)\big)$, are individual ethical dimensions of world states that must be avoided in order to meet ethical constraints. *Risk factors* are features whose quantitative assignment is to be minimized in order to meet ethical expectations. *Ethical opportunities* correspond to aspects world states that must be favoured in order to meet ethical constraints. *Opportunity level factors*, denoted by $\vec{e}^{\,o}(i) = \big(e^o_1(i), \ldots, e^o_k(i)\big)$, are features (e.g. GENDER EQUALITY) whose quantitative assignment is to be maximized in order to meet ethical expectations.

Ethical induction depends on how risks and opportunities contribute to the overall *ethical signature* $\vec{es}(i)$ of an individual case i. The training data set T includes a reference (gold) feature vector $\vec{i} = \big(\vec{f}(i) \,||\, \vec{es}(i)\big)$ that concatenates the original evidences $\vec{f}(i)$ with $\vec{es}(i) = \big(\vec{e}^{\,r}(i) \,||\, \vec{e}^{\,o}(i)\big)$ expressing all the ethical implications of $\mathcal{EO}$ against the decision $d(i)$. The enriched training instances form the overall *ethically enriched training set* T^{eth}, defined as:

$$T_{eth} = \Big\{\big(\vec{i}, d(i)\big) | i \in T\Big\} = \Big\{\Big(\big(\vec{f}(i) \,||\, \vec{es}(i)\big), d(i)\Big) | i \in T\Big\}$$

that can suitably support multitask, i.e. business *and* ethical, learning.

Notice that the *Ethical status* of an instance i can be derived as a function of the $\vec{es}(i)$ vector: ethical states are a discrete set of categories defined by thresholding over risks and opportunity distributions. In order to synthesize the ethical description of an instance, the overall benefit and risk of an instance form a pair of stochastic variables $(\mathcal{B}, \mathcal{R})$ whose values are derived from the probability distributions of individual opportunity levels (e^o_j) and risk factors (e^r_k), respectively. In future, trained systems are expected not to promote/suggest decisions $d(i)$ that result in an ethical status of future instances i that is *not less than mildly ethical*. This graded judgment will be made dependent on the $(\mathcal{B}, \mathcal{R})$ states derived from the probability distributions in the signature $\vec{es}(i)$.

Ethical Features and Inductive Reasoning

Risk factors and Opportunity levels, described by $\vec{es}(i)$, express how individual observable features $f_j(.)$ trigger ethical aspects. Truth-makers in the ethical ontology $\mathcal{EO}$ act on observable features $f_j(i)$ (e.g., SEX = "$female$") and determine corresponding values onto ethical features (e.g. the es_j that represents the j-th ethical dimension). These assignments are determined by complex reasoning chains possibly depending on multiple features or multiple instances. Individual risks and opportunities correspond to dimensions that can be multiply assigned by different truth-makers.

Probabilistic restrictions over the domains of risks and opportunities allow to vectorially represent the ethical signature. Whenever an instance $i \in T^{eh}$ activates one or more rules in $\mathcal{EO}$, the truth-makers set the corresponding k-th ethical opportunity or risk factor $es_k(i)$ to the predicted status of the k-th ethical dimension. Multiple rules may affect the same ethical factor and a cumulative effect is obtained. We thus model the ethical signature vector with as many values as they are foreseen in the corresponding domain of a risk and opportunity factor: if B is the number of opportunities, R is the number of risks and V is the number of values in their domains, the overall number of ethical risk and opportunity dimensions is $(B + R) \cdot V$.

A pair *instance-decision* implies ethical consequences, i.e., *ethical risks* and *ethical opportunities*, that are not hardcut. They can be captured by graded judgments along the ethical dimensions, e.g., probability distributions over the

reference domain. While other design choices are in principle possible, we propose to discretize every ethical dimension in the same domain $\mathcal{V}$ defined by a finite, closed and ordered set: $\mathcal{V} = \{v_i \in \mathbb{R} : 0 \leq v_1 < ... < v_m \leq 1\}$. In particular, for both benefits and risks, we fixed $m = 5$ and limit values in the $[0, 1]$ range. The following five labels can be adopted {"VERY LOW", "LOW", "MILD", "HIGH", "VERY HIGH"} corresponding to the numerical values $v_1 = 0.1, v_2 = 0.25, v_3 = 0.5, v_4 = 0.75$ and $v_5 = 0.9$.

The role of truth-makers. Truth-makers are the rules of the $\mathcal{EO}$ ontology that actively determine the ethical profile of the instance-decision $(i, d(i))$ pair. In particular, given a pair $\big(i, d(i)\big)$, a truth-maker tm will determine a probability distribution to the set of benefit and risk dimensions. For every tm, ethical dimension $e_j(i)$ and possible ethical value $v_k \in V$ the following probability is defined:

$$P\Big(e_j(i) = v_k \mid \big(\vec{i}, d(i)\big), tm\Big) \qquad \forall j, \forall k = 1, \ldots, 5$$

which expresses the evaluation of the truth-maker tm onto the instance i given the decision $d(i)$, along the k-th value of the j-th ethical dimensions. A truth-maker thus assigns probabilities to the ethical signature of an individual i for all possible combinations of business characteristics $\vec{f}(i)$ and decisions $d(i)$[1]. Multiple truth-makers can contribute to a given ethical feature $e_j(i)$ individually biasing their overall probability $P(e_j(i))$. When all truth-makers are fired, the resulting *ethical signature* over an instance $\vec{i}$ and its decision $d(i)$ consists $\forall j, k$:

$$es_j(i) = \prod_{tm} P\Big(tm|\mathcal{EO}\Big) P\Big(e_j(i) = v_k \mid \big(\vec{i}, d(i)\big), tm\Big)$$

Notice that when a training instance is defined, the unique decision $d(i)$ is available and one unique ethical signature is the result. During classification no final $d(i)$ is available and the estimates of the ethical implication must be available for all the different target classes, $d_1, ..., d_l$. From signatures we can then express the final ethical status. Notice also that individual decisions over input i correspond to probabilities along all the dimensions determined by decisions, risks and opportunities. A factor y^{ljk} estimates the probability of the *joint* event $\big(d(i), \mathcal{B}, \mathcal{R}\big)$ corresponding to i. By assuming independence, each element y^{ljk} estimates the following:

$$\begin{aligned} y^{ljk} &= P\big(d(i) = d_l\big) \cdot P\big(\mathcal{B} = v_j\big) \cdot P\big(\mathcal{R} = v_k\big) \\ &= \text{(shortened as) } P(d_l) \cdot \mathcal{B}_j \cdot \mathcal{R}_k \end{aligned} \tag{1}$$

The collective benefit $\mathcal{B}$ is obtained as a joint probability distribution:

$$\mathcal{B}_j = P\big(\mathcal{B} = v_j\big) = \prod_{t=1}^{B} P\big(e_t^o(i) = v_j|\vec{i}, d(i)\big) P\big(e_t^o(i)\big)$$

where $P\big(e_t^o(i)\big)$ is the probability of the t-th ethical feature in describing the collective benefit $\mathcal{B}$. Similarly, risk $\mathcal{R}$ is modeled as joint probability distribution whose component are defined by:

$$\mathcal{R}_k = P(\mathcal{R} = v_k) = \prod_{t=1}^{R} P\big(e_t^r(i) = v_k|\vec{i}, d(i)\big) P\big(e_t^r(i)\big)$$

Variables y^{ljk} control the impact of risks and opportunities during training and can be assigned to specific neurons.

Gold standard for Ethics: Ethical landmarks. Given the ethical signature $\vec{es}(i)$ of an instance i, we can reason about its ethicality. Two specific points in the ethical domain can be defined as references for a quantitative measure of ethical sustainability and unacceptability. The probability mass function reserving most of the probability to v_5 ="VERY HIGH" to ethical benefits while minimizing the probability of ethical risks v_1 ="VERY LOW" is by definition the *ethical optimum* (OPT_{eth}). Dually, we define the *ethical minimum* (MIN_{eth}) as the probability distribution that reserves most probability to the minimum opportunity value, v_1 and maximal probability to the maximal risk, v_5. During training, every ethical signature is optimized to be close to the ethically optimal and far from the ethical minimum.

DEFINITION: (*Ethical compliance*). An instance-decision pair $(\vec{i}, d(i))$ is *ethically compliant to* $\mathcal{EO}$ iff:

$$dist\big(\vec{es}(i, d), MIN_{eth}\big) \geq dist\big(\vec{es}(i, d), OPT_{eth}\big)$$

where $\vec{es}(i, d)$ is the ethical signature of i given the decision d and $dist$ is a valid distance over probability distributions.

Embedding Ethics as Multitask Neural Learning

Once a quantitative model for ethics is available through observable features, risk factors and opportunity levels as probability distributions across finite domain $\mathcal{V}$, neural learning is enabled. An ethical neural architecture should be able to use dependencies among observable features as triggers of the target business decisions but also to actively recognize dependencies between ethical and observable features, i.e. ethical consequences implied by some features.

In this perspective, back-propagation has the aim of optimizing both the business accuracy and the ethical compliance. For this reason, we propose the adoption of a multi-strategy learning approach with the cascading (i.e. stacking) of different (sub)networks. The proposed network is composed of 3 main processing stages, as shown in Figure 1. In the first stage the input vector $\vec{x}$ is fed to a series of fully connected layers, namely the **Ethics encoder**. Its role is to learn combinations of input features able to capture relationship between business observations and, possibly, their ethical consequences (i.e., ethical features). Later stages of the network can exploit the effective ethics encoding without resorting back to the $\mathcal{EO}$. This component is not directly optimized through a loss function but, rather, it receives penalties by back-propagation from later layers. It can be seen as a sort of pre-training stage. Formally, it corresponds to: $\Phi(\vec{x}) = g_1(W_1\vec{x} + \vec{\theta}_1) = \hat{y}_1 \in \mathbb{R}^{d_1}$ where $W_1 \in \mathbb{R}^{(n+2mj)\times d_1}$, $\vec{\theta}_1 \in \mathbb{R}^{n+2mj}$ are parameters to be optimized, d_1 is a network meta-parameter. The second stage

[1] If no truth-maker is triggered by an instance the uniform probability distribution u is used, i.e. $P\Big(e_j(i) = v_k|\big(\vec{i}, d(i)\big), tm\Big) = \frac{1}{m}$, over the values v_k and different ethical features, i.e. $\forall j, k$.

comprises two MLPs that are independently trained to learn two different tasks: estimating the correct decisions' distribution, under the sole business perspective, and to reconstruct the ethical consequences of such decisions. The Business Expert DNN and the Ethics Expert DNN are responsible for the first and the second task, respectively. Note that they receive the same input, that is the vector emitted from the first stage of the architecture.

The Business Expert (BE) DNN. As it's entrusted with emitting business decisions without any direct penalization for the unsatisfactory ethical consequences, it can be seen as the final layers of an ethics-agnostic sub-network, modeled as $BE\big(\Phi(\vec{x})\big) = BE\big(\hat{\vec{y}}_1\big) = g_2\big(W_2\hat{\vec{y}}_1 + \vec{\theta}_2\big) = \hat{\vec{y}}_2 \in \mathbb{R}^K$ where K is the number of output categories, i.e. decisions. The estimator is then optimized by a standard cross-entropy loss function over the predicted distribution $\hat{\vec{y}}_2$ against the gold distribution $d(i) = \vec{y}_B$.

The Ethical Expert (EE) DNN. Its role is to reconstruct the ethical signature for each pair (x_{input}, d). It processes the encoding from the first stage and it outputs a vector which represents the *joint* probability of the triplet $(decision, benefit, risk)$ under maximal entropy of the business decisions distribution and independence assumption. Here, the EE is modelled as $EE(\Phi(\vec{x})) = EE(\hat{\vec{y}}_1) = g_3(W_3\hat{\vec{y}}_1 + \vec{\theta}_3) = \hat{\vec{y}}_3 \in \mathbb{R}^{K\times m^2}$ where K is the number of possible decision and m is the number of possible values for ethical benefits and risks. As in Equation 1, each element $\hat{y}_3^{ijk}$ in the output vector should reconstruct $y_3^{ijk} = u_d \cdot P(e_j^o) \cdot P(e_k^r)$ where u_d is the expected value of the uniform distribution over the possible decisions and the probabilities for benefits and risks are the ones in the corresponding ethical signature. Then, the cross-entropy loss function $\mathcal{L}_{Er}$ is applied to compute the ethics recognition loss over the predicted $\hat{\vec{y}}_3$ against the gold distribution encoded in the vector $\vec{y}_3$.

Ethics-aware (EA) Deep Neural Network. Similarly to the EE network, it is responsible for estimating the joint probability of each possible triplet (d_i, b_j, r_k). However, here $P(D = d_i)$ is directly derived from the gold standard while the probabilities for benefits and risks are extracted from OPT_{eth} for ethically compliant decisions and MIN_{eth} for not compliant ones, i.e.:

$$y_5^{ijk} = P(d_i) \cdot P(b_j^{opt}) \cdot P(r_k^{opt}) \quad (\text{if } (\vec{x}, d_i) \in \mathcal{D}^+)$$
$$y_5^{ijk} = P(d_i) \cdot P(b_j^{min}) \cdot P(r_k^{min}) \quad (\text{if } (\vec{x}, d_i) \in \mathcal{D}^-)$$

where $\mathcal{D}^+$ and $\mathcal{D}^-$ are the set of ethically and not compliant decisions for $\vec{x}$, respectively, according to $\mathcal{EO}$. Overall, this sub-network is described by

$$EA([\hat{\vec{y}}_2; \hat{\vec{y}}_3]) = EA(\hat{\vec{y}}_4) = g_4(W_4\hat{\vec{y}}_4 + \theta_4) = \hat{\vec{y}}_5 \in \mathbb{R}^{K\times m^2}$$

At this stage, as for the Ethics Expert, the error is updated by computing the Ethical Loss $\mathcal{L}_E$, which is again the cross entropy between y_5 and $\hat{\vec{y}}_5$. Note that this formulation is not directly promoting ethically sustainable decisions but it is rather encouraging the network to pair them with highly-beneficial and low-risk ethical consequences.

The final business decision of our network is determined by a *decision policy* over risks and opportunities. Here we define two possible policies:

- **Ethics-Unconstrained (EU) policy.** The final decision d_i is derived simply by summing up all probability contributions of the triplets (i, j, k) where i is fixed, i.e.,

$$\hat{d}_i^{EU} = argmax_i P_{EU}(d_i) \tag{2}$$

- **Ethics-Constrained (EC) policy.** Here a probability $P(d_i, be_j, ri_k)$ contributes to $P(d_i)$ only if be_j, ri_k satisfy some membership constraints, i.e.,

$$\begin{aligned}\hat{d}_i^{EC} &= argmax_i P_{EC}(d_i) \\ &= argmax_i \sum_{v_j \in \mathcal{V}'} \sum_{v_k \in \mathcal{V}''} P(d_i, v_j, v_k)\end{aligned} \tag{3}$$

  where we set $\mathcal{V}' = \{\text{“}\texttt{HIGH}\text{”}, \text{“}\texttt{VERY HIGH}\text{”}\}$ and $\mathcal{V}'' = \{\text{“}\texttt{LOW}\text{”}, \text{“}\texttt{VERY LOW}\text{”}\}$.

As we will see in the experimental evaluation, the above network is able to learn from a business point of view (through the loss $\mathcal{L}_F$) consistently with the $\mathcal{EO}$ (through the ethical loss $\mathcal{L}_E$), while promoting ethically sustainable business decisions.

Balancing business and ethical adequacy

Different contexts and applications may require different trade-offs between the prescriptions from the ethics system and the behavioural patterns induced from historical data. While it may be possible to balance such trade-offs by changing the distributions of benefits and risks derived from the truth-makers, it would be cumbersome from a practical point of view and, more importantly, it may make it difficult to compare different models, as the underlying feature spaces would be related but different. More manageable methods, as discussed here, consist in acting only on the joint distributions used to train the EA-DNN.

Smoothing business decision. Gold standards usually provide unique decisions, that is sharp probability distributions across business decisions. However, they are not guaranteed to be ethical. Smoothing is needed to allow the neglecting of unethical cases so that probabilities for decisions d_i different from the gold standard decisions are non zero. *Laplace smoothing* is applied across the K different classes, expressed by: $\hat{d}_i = \frac{d_i + \alpha}{1 + K\alpha}$.

Tweaking between Ethics and Business accuracy. Through similar a technique, it's possible to tune the emphasis of ethical consequences by applying a *tweaking factor* β to the probability of benefits and risks in the joint probability of the $(d, \mathcal{B}, \mathcal{R})$ triple, i.e.,:

$$P_\beta\big(d_i, \mathcal{B}_j, \mathcal{R}_k\big) = P\big(d(i) = d_l\big) \cdot \big(P(\mathcal{B} = v_j) \cdot P(\mathcal{R} = v_k)\big)^\beta$$

Here the influence of ethics turns weaker as $\beta \to 0$. The above equation corresponds to the input of the network and establishes the influence onto the NN of the ethical information through the corresponding impact on loss functions $\mathcal{L}_{Er}$ and $\mathcal{L}_E$.

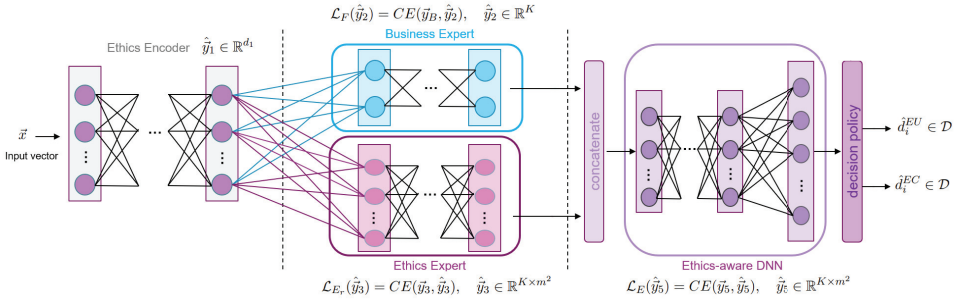

Figure 1: The architecture of the Ethical by Design Neural Network.

Experimental Evaluation: Ethical Risk Assessment in Banking

We run extensive evaluation of the proposed framework on the German Credit dataset[2] (GC). Here the task is to predict whether a loan request carries a " low" ($C0$) or " high" ($C1$) risk of default (i.e., the requester not paying back the loan) based on 20 different attributes, some of which are domain-specific (e.g., PREVIOUS CREDIT HISTORY or ACCOUNT BALANCE) and others are more general (e.g., AGE of the requester or the NUMBOFPEOPLEUNDERMAINTENANCE). Despite its small dimension (only 1000 instances) and strong class unbalance (700 instances labeled as "low" $C0$ profile), this dataset is appealing to test ethics learning approaches as it represents a real-world problem (King, Feng, and Sutherland 1995) and offers many attributes upon which ethical rules can be defined. We defined and experimented on different ethical policies and truth-makers but, due to space limitations, we will focus on one particularly simple ethic ontology ($\mathcal{EO}_1$), which includes two truth-makers: "MOTHERHOOD FOSTERING" (tm_{MF}), favouring (lending decisions representing) women with children and, to a lesser extent, men with at least 2 children, and "CULTURAL INCLUSIVENESS" (tm_{CI}), favouring foreign workers. Details on the truth-makers are in Appendix[3]. Ethical values $V = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ are used to express $\mathcal{V} = \{$"VERY LOW", "LOW", "MILD", "HIGH", "VERY HIGH"$\}$. Due to the strong unbalance between the target classes (70%-30%), we report business performances according to the average F1-measure, $\mu F1$, as: $\mu F1 = \frac{F1_{C0}+F1_{C1}}{2}$. The overall *ethical compliance* $EComp$ of the data set, given the ontology $\mathcal{EO}_1$, is computed as the percentage of ethically complaint instances, according to the gold standard decision, i.e. $\frac{\mathcal{D}^+}{\mathcal{D}^+ + \mathcal{D}^-}$. It corresponds to the $EComp = 0.70$ that suggests that historical data alone cannot be used to promote ethics.

It is clear how the joint adherence to the $\mathcal{EO}_1$ ethics and to business optimality requires a complex trade-off. It requires in fact neglecting some training cases to improve upon ethical compliance. A possible straightforward measure of the trade-off between ethics and business accuracy is the parametrized *Acceptability factor* $EAcc_\gamma$ as the weighted average between the $\mu F1$ and the ethical compliance $EComp$:

$$EAcc_\gamma = \gamma \cdot \mu F1 + (1-\gamma) \cdot EComp \qquad (4)$$

where $\gamma \in [0, 1]$ can be adjusted according to the relative importance of the two terms. The $EAcc_\gamma$ measure, when the superiority of ethics is imposed by $\gamma = 0.2$ over the GC dataset, provides the strong baseline for ethical training given by $EAcc_{.2} = 0.76$. Such gold standard $EAcc_\gamma$ is a useful reference measure to compare ethical neural models.

Experimental Set-Up. The chosen architecture for the EbDNN has an Ethics Encoder with 2 layers, where the first layer has the same size of the input and the second has dimension 400, the Business Expert has 1 layer with output dimension equals to K. Both the Ethics Expert and the Ethics-Aware DNN have 1 layer with $K \cdot m^2$ neurons (where K is the number of classes and m the number of ethical values). Non-linearity is applied through the *relu* function at each layer, except for the last layer in each component associated with a loss function, where a *softmax* is computed. A dropout rate of 0.2 on each layer is applied. To cope with the limited number of instances, we applied 10-fold cross validation, training each model for 1000 epochs with a standard batch size of 256 through Adam optimizer[4]. Various settings of the the smoothing and tweaking factors $(\alpha, \beta) \in \{0.1, 0.3, 0.6, 1.0\} \times \{0.01, 0.05, 0.10, 0.20, 0.35, 0.50, 0.75, 1.00\}$ have been applied to systematically study their impact. We fed each model alternatively with the enriched input vector, i.e., $\left(\vec{f}(i) \,||\, \vec{es}(i)\right)$ or only with business observable $\vec{f}(i)$. No significant difference has been observed as the EE-DNN seems able to robustly reconstruct ethical signatures across all settings. In the rest of the experiments, we thus trained models only over $\vec{f}(i)$.

To provide a fair comparison with a standard learning framework, a simple MultiLayer Perceptron (MLP) with 2 layers and 320 units per layer has been trained, over $\vec{f}(i)$ only: it achieves an accuracy of 76.21% comparable to state-of-the-art results on this dataset (Ratanamahatana and Gunopulos 2002). It corresponds to a business performance of $\mu F1 = 66.4\%$ with an ethical compliance $EComp = 77.8\%$. The ethical acceptance is thus

[2]Publicly available from the University of California-Irvine machine learning repository (Dua and Graff 2017).

[3]Supplemental material in the submitted version.

[4]Models were implemented in python using Tensorflow.

System (α, β)	$\mu F1$	$EthCompl$	$EAcc_{0.2}$
$EA^{EU}(0.3, 0.5)$	63.1%	79.6%	76.3%
$EA^{EU}(0.3, 0.01)$	63.9%	78.8%	75.8%
$EA^{EC}(0.1, 0.5)$	41.2%	100.0%	88.2%
$EA^{EC}(0.1, 0.2)$	53.8%	93.0%	85.2%
$EA^{EC}(0.1, 0.01)$	61.7%	78.4%	75.1%
$EA^{EC}(0.3, 0.1)$	60.6%	85.1%	80.2%
MLP	66.4%	77.8%	75.5%

Table 1: $\mu F1$, $EthCompl$ and $EAcc_\gamma$ ($\gamma = 0.2$) for different configurations of the EA model.

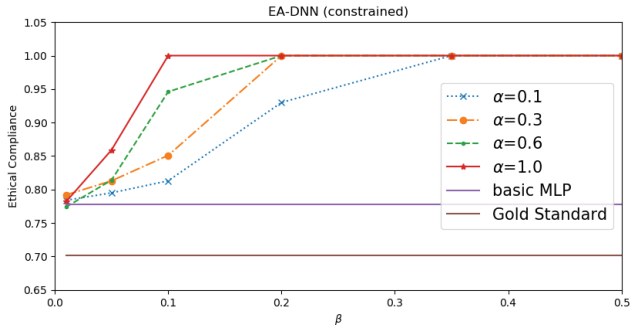

Figure 2: The trends of the Ethical Compliance $EComp$ of the outcome of the EA-DNN as a function of the tweeking β. While MLP and Gold Standard refers to ethically unaware methods, plots represent several smoothing α parameters.

$EAcc_{.2} = 75.5\%$ that does not improve on the gold standard, as expected: it provides a second comparative reference as ethical unaware system.

Evaluating ethical aware learning. Table 1 reports the performances of both the baseline MLP and of the EA models, under different α,β settings and decision policies. The trade-offs between ethical and business performances is largely improved by EA models for all the configurations. Gains in ethical compliance of EA models w.r.t. baselines are significant while business performance losses are relatively small.

The effect of both factors (α, β) is observed in Figure 2. As β increases, ethics plays a stronger role and the model's behaviour deviates from a purely business-driven predictor. The smoothing factor α plays a complementary role: stronger smoothing actions corresponds to markedly more ethical behaviours, even for smaller β. Notice how, even for high α values at lower β's ($\leq$ 0.1) every EA models starts to exhibit unethical choices. The fully enforced ethics network EA^{EC} with $(\alpha, \beta) = (.1, 0.5)$ achieves the maximal $EthCompl$ with less than 20% loss in terms of $\mu F1$. Note that, the *unconstrained* decision policy, i.e., the EA^{EU} model, is not sensitive to the tweaking factor, as for $\beta = 0.5, 0.2$ the performance is basically the same. Figure 3 plots Ethical Acceptability $EAcc_\gamma$ (with $\gamma = 0.2$) restricted to the test cases where the MLP provides non ethical decisions. The robustness of ethical aware networks is striking. The progressive deviation from ethical sustainability is visually captured in Fig 4, where the *ethical signature* of each

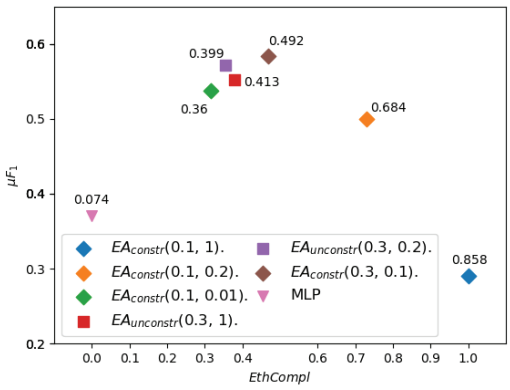

Figure 3: $EAcc_\gamma$ for $\gamma = 0.2$ of the EA model over non ethical decisions of the gold standard: performances of constrained (EA^{EC}), unconstrained (EA^{EU}) networks and the baseline MLP are reported against $EthCompl$ and μF_1 values.

prediction is projected on the plane with axis $(\mathbb{E}[be], \mathbb{E}[ri])$ (the point size is proportional to the number of projections falling in that point): as β decreases, more and more points are mapped in the semi-plane where the expected value of risk is higher than the expected benefit.

Overall, the experimental evaluation confirmed that the embedding of ethics principles into the decision function of the model can be effectively modulated through the fine tuning of β, and to lesser extent α, and the application of the proper decision policy.

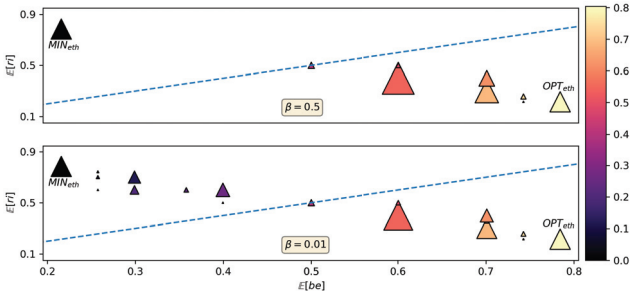

Figure 4: Projections of ethically-constrained EA-DNN's predictions for different $\beta = 0.5$ (top) and 0.01 (bottom) values: size is proportional to the number of projected predictions. The blue dashed line corresponds to ethically neutral choices, i.e., expectation about benefits is equal to risks, hence the upper left half-plane includes all ethically non compliant decisions.

Conclusions

In this work, we propose a deep learning framework to achieve the acquisition of high quality inferences that simultaneously reflects ethical expectations. Experimental evaluation suggests the framework to be effective as well as to allow the fine-tuning of the balance between business

and ethics perspective, through the smoothing and tweaking methods. This work represents an early exploration of the framework potential, hence future directions are rich as they range from the definition of more complex ethics and the application to more challenging inference tasks as well as different learning paradigm (e.g., Reinforcement Learning).

Appendix A: a Simple Ethical Ontology for the Lending Case

During experimental evaluation, we tested the learning framework against a very essential ethics $\mathcal{EO}_1$, enforced by 2 truth-makers: 'MOTHERHOOD FOSTERING' (tm_{MF}) and 'CULTURAL INCLUSIVENESS' (tm_{CI}), both promoting the 'low risk-profile' decision as ethically preferable when certain conditions are met. If the instance description doesn't satisfy the conditions of any rule, then the default behaviour is to assign ($:=$) probability distributions centered in $(\mathcal{B} := \texttt{MILD}, \mathcal{R} := \texttt{MILD})$.

In the following, req stands for the loan request, $(\mathcal{B} := v_i, \mathcal{R} := v_j)$ indicates the assignments of gaussian distributions centered in v_i and v_j to benefits $\mathcal{B}$ and risks $\mathcal{R}$, respectively. Due to limited space, here we report the details of tm_{MF} only:

'MOTHERHOOD FOSTERING':

$sex(req, female) \wedge maintainedPeople(req, X) \wedge X \geq 2$
$\wedge \quad loanRisk(req, low)$
$\Rightarrow (\mathcal{B} := \texttt{VERY\_HIGH}), (\mathcal{R} := \texttt{VERY\_LOW})$

$sex(req, female) \wedge maintainedPeople(req, 1)$
$\wedge \quad loanRisk(req, low)$
$\Rightarrow (\mathcal{B} := \texttt{HIGH}, \mathcal{R} := \texttt{LOW})$

$\neg sex(req, female) \wedge maintainedPeople(req, X) \wedge X \geq 2$
$\wedge \quad loanRisk(loan, low)$
$\Rightarrow (\mathcal{B} := \texttt{HIGH}, \mathcal{R} := \texttt{MILD})$

$sex(req, female) \wedge maintainedPeople(req, X) \wedge X \geq 2$
$\wedge \quad loanRisk(req, high)$
$\Rightarrow (\mathcal{B} := \texttt{VERY\_LOW}, \mathcal{R} := \texttt{VERY\_HIGH})$

$sex(req, female) \wedge maintainedPeople(req, 1)$
$\wedge \quad loanRisk(req, high)$
$\Rightarrow (\mathcal{B} := LOW, \mathcal{R} := \texttt{HIGH})$

$\neg sex(req, female) \wedge maintainedPeople(req, X) \wedge X \geq 2$
$\wedge \quad loanRisk(req, high)$
$\Rightarrow (\mathcal{B} := \texttt{LOW}, \mathcal{R} := \texttt{MILD})$

Consider, for examples, the two instances partially represented in Table 2,i.e. a case (1) representing a female requester with 1 child and associated with an high risk profile and a case (2) representing a male requested with 2 children and associated with a low risk profile. Then, ethical signatures $\vec{es}_i$ that are derived from the triggering of rule (6) and (5) in tm_{MF}, respectively, will promote the decision regarding (2) as beneficial (i.e., high ethical benefit and low ethical risk), while they will penalize with similar intensity the decision regarding case (1).

REQ	SEX	MANTAINEDP.	LOANRISK
1	*female*	1	*high* $C1$
2	*male*	2	*low* $C0$

Table 2: Two (partial) examples of instances in the German Credit dataset.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature* 563.

Bennett, M. 2013. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation* 14(3-4):255–268.

Bird, S.; Barocas, S.; Crawford, K.; and Wallach, H. 2016. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT-ML), New York University*, 4.

Boddington, P. 2017. *Towards a Code of Ethics for Artificial Intelligence*. Springer.

Bonnemains, V.; Saurel, C.; and Tessier, C. 2018. Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*.

Bostrom, N., and Yudkowsky, E. 2014. *The ethics of artificial intelligence*. Cambridge University Press. 316–334.

Craglia, M. 2018. *Artificial Intelligence: A European Perspective*. Publications Office of the European Union.

Dua, D., and Graff, C. 2017. UCI machine learning repository.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

King, R. D.; Feng, C.; and Sutherland, A. 1995. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal* 9(3):289–333.

Kleiman-Weiner, M.; Saxe, R.; and Tenenbaum, J. B. 2017. Learning a commonsense moral theory. *Cognition* 167:107 – 123. Moral Learning.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Crown Publishing Group.

Ratanamahatana, C. A., and Gunopulos, D. 2002. Scaling up the naive bayesian classifier: Using decision trees for feature selection. In *Workshop Data Cleaning and Preprocessing (DCAP 2002), at IEEE Int'l Conf. Data Mining, ICDM 2002*. Citeseer.

Smuha, N. 2019. *Ethics guidelines for trustworthy AI*. Publications Office of the European Union.

Vanderelst, D., and Winfield, A. 2018. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research* 48:56 – 66. Cognitive Architectures for Artificial Minds.