ETHICS OF AI

GROUP - 3

Ethico-legal Governance: AI in Employment

Authors:

HELENA WINIGER

RAHAF GHARZ ADDIEN

AMARDEEP PAL

FEDERICA SURIANO

MUHTASHIM LEKHON

Submitted: August 1, 2021

Introduction

This project is a study on the idea of ethico-legal governance based on symbolic AI. The aim of the project was to answer its research question on the possibility of encoding ethical and legal constraints in symbolic logics, to explore the required ethics and symbolic logics for this task as well as the needed governance architectures.

As a case study, we are interested in the use of AI in hiring that was classified as high risk in the EU proposal for a regulation of AI (AI Act)[1] which emphasizes the need for governing such AI systems ethically and legally. We tried to examine possible levels of legal governance based on the related articles of General Data Protection Regulation (GDPR)[2] and encode the articles using symbolic and deontic logics.

In the first section, the idea of ethico-legal governance and its importance will be presented and, thus, the risk of using AI in employment will be discussed. Moreover, the results of our project's implementation after explaining its principle concepts, namely symbolic AI, Benzmüller et al.'s LogiKEy[3] as well as the according deontic logics will be introduced as the reasoning process. Reflection on the interdisciplinary teamwork is delineated in the second section while in the third section, the project will be reflected in a broader ethical context, including a technical, social and research perspective.

1 Description of the project

1.1 The principle of ethico-legal governance

The principle of ethico-legal governance, here, is referring to an enforceable concept that ensures machine learning (ML) technologies are used only for human benefits: It is, therefore, human-centric instead of instrumentalized for specific self-interests of parties involved. Ideally, ethico-legal governance is integrated as a component in the architecture of AI systems by the developer[4], and required in the work of numerous governments and policymakers. The dilemma between human rights and infringement always creates a gap between accountability and ethics in technological advancement. Ethico-legal governance is a first step to approach closing the gap in serving the common good. The wholesome idea behind this form of AI governance is to establish justice, data quality, and maintaining the autonomy of the individual as the subject of fundamental rights. This requires a specific and transparent solution regulating AI.

As shown by Theodorou[5], ML is sometimes biased, e.g., in unfairly rejecting individuals for loans, inaccurately recognizing basic information about users, and particularly in racial profiling. Because of an algorithm's behavior, Lufthansa's costs for local flights within Germany increased up to 30% after Air Berlin called off bankruptcy. After the investigation, the authority of Lufthansa claimed that the company has no control over the algorithm's autonomous behavior[5]. Many experts are concerned with the unrestricted use of such 'black box' systems in e.g., finance, education, criminal justice, search engines, or even in social welfare.

Cath points out that ML and algorithmic operations create many socio-technical challenges[6], especially in combination with regulations like the GDPR. However, European data protection provides a robust frame that results in a slow development of interpretations and justified concerns.

In summary, when AI systems, e.g., with ML algorithms, are involved in making decisions which are including black box elements, AI governance is a must to gain reason based decisions and observe human rights and values: e.g., in preventing bias and ensuring diversity, non-discrimination, and fairness[7,8].

1.2 Symbolic AI

AI is a discipline that was born in the 1950s[9], a period of great scientific ferment on the study of the computer and its use for intelligent systems. The initial idea of AI was based on optimism in problem solving and reasoning. The dominant paradigm was symbolic AI until 1980[9].

Symbolic AI is a reasoning oriented sub-field of AI that focuses on research based on high-level symbolic (human-readable) representations of problems and logics[10]. After the 1980s, however, subsymbolic AI has been starting taking the lead and gaining attention until the recent years[9].

Symbolic techniques are defined by explicit symbolic methods, such as formal methods and programming languages, and are usually used for deductive reasoning[9] in order to produce logical conclusions. In addition to the ability to explain the reached conclusion, symbolic methods are able to explain their intermediate steps due to the modularity[11] of which they are characterized. These steps are based on rules which set up discrete and autonomous units of knowledge that can be easily inserted or removed from a knowledge base.

In this project, ethico-legal governance aims basically at justifying the decisions made using symbolic AI techniques, thanks to which we are able to translate some formally codified ethico-legal theories into actions[12].

1.3 AI in employment as high risk systems

According to the European Commission's (EC) proposed AI Act, the use of trustworthy AI is classified in a risk-based spectrum, namely unacceptable risk, high risk, limited risk, and low risk. High risk is determined by the potential and probability of harming individuals. "Harm" can be understood as a high risk to human health and safety or to fundamental human rights. According to the proposal, a high risk AI system shall not only be limited by the functionality of the system but also by the specific purpose and modalities for which it is used.

Therefore, the use of AI in employment and workers management is listed in the proposal's high-risk category. In this case, algorithms used for drawing conclusions influence the decision-making process regarding individuals and could eventually impact the future career prospects and livelihoods of these individuals negatively. Such decisions may endorse patterns that have a history of discrimination, for example against women, certain age groups, or people of certain racial or ethnicity or sexual orientation[1]. For instance, Amazon used AI tools for its hiring process, which in 2018, found to be discriminatory against women[13]. Furthermore, AI used to monitor performance and behavior may also violate privacy and data protection rights[14].

1.4 The reasoning process

In this part, we will dive deeper into how legal governance could be implemented in practice using symbolic AI. The implementation follows and depends basically on the methodology and framework of LogiKEy presented by Benzmüller et al.[3,15] for the designing and engineering of ethical reasoners, normative theories, and deontic logics. It enables assessing the suggested actions before executing them based on the modeled ethico-legal theories and ensures high degrees of reliability, transparency, and explainability.

Figure 1 illustrates the structure of the explicit ethical reasoner where the input is actions to be decided on according to the ethico-legal and domain theories, and which are suggested by the AI reasoner based on the relevant application data, whereas the output is the accepted actions and the corresponding explanations. Therefore, we need to identify the relevant legal theories to be encoded, the suitable level of abstraction, and the appropriate deontic logic.

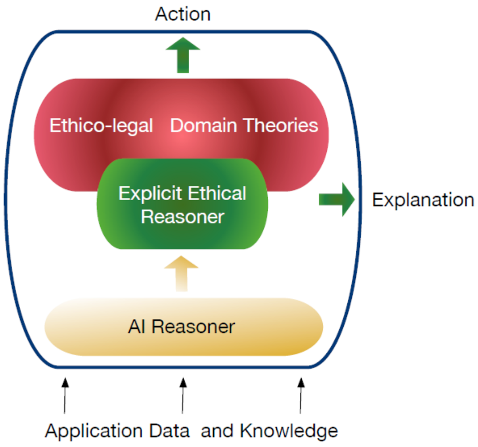
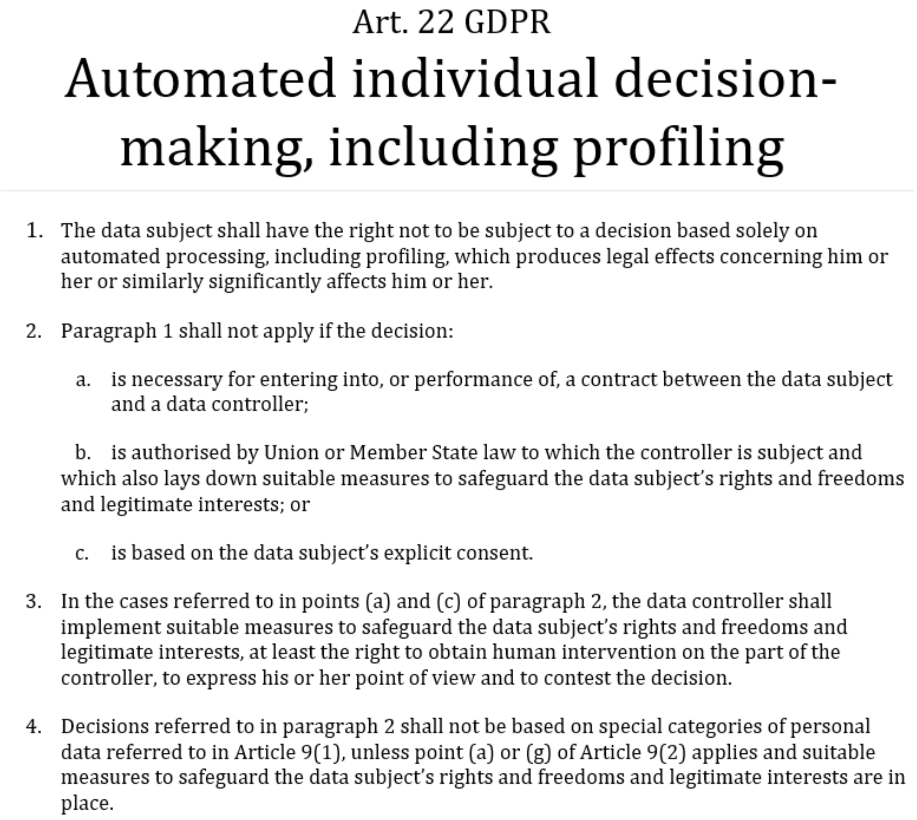


Figure 1. Explicit ethical reasoner for intelligent autonomous systems,[3]

1.4.1 Legal theories

Our approach for governing AIs in employment focuses on GDPR[2] as a legal theory, namely Article. 22, Figure 2, which determines the cases where automated individual decision making can be applied and the measures that have to be considered in each case. According to Article. 22 of GDPR and regarding the case of the hiring problem, the data subject is allowed to be subject to a decision based solely on automated processing since the decision is necessary for entering into a contract between the data subject and the data controller. However, suitable measures to safeguard the data subject's rights, freedoms, and legitimate interests have to be applied, at least human interventions to evaluate the decision.

Article. 22 refers also to Article. 9, Figure 3, which defines special categories of the personal data based on which the automated decisions should not be made unless it is necessary for reasons of substantial public interest - and suitable measures have to be applied here as well. In its wider context, the here handled use case of AI in employment is legally embedded in the definition of High Risk[1], and based on the EU Charter of Fundamental Rights[16] which is also legally binding for GDPR.



Art. 22 GDPR

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

   a. is necessary for entering into, or performance of, a contract between the data subject and a data controller;

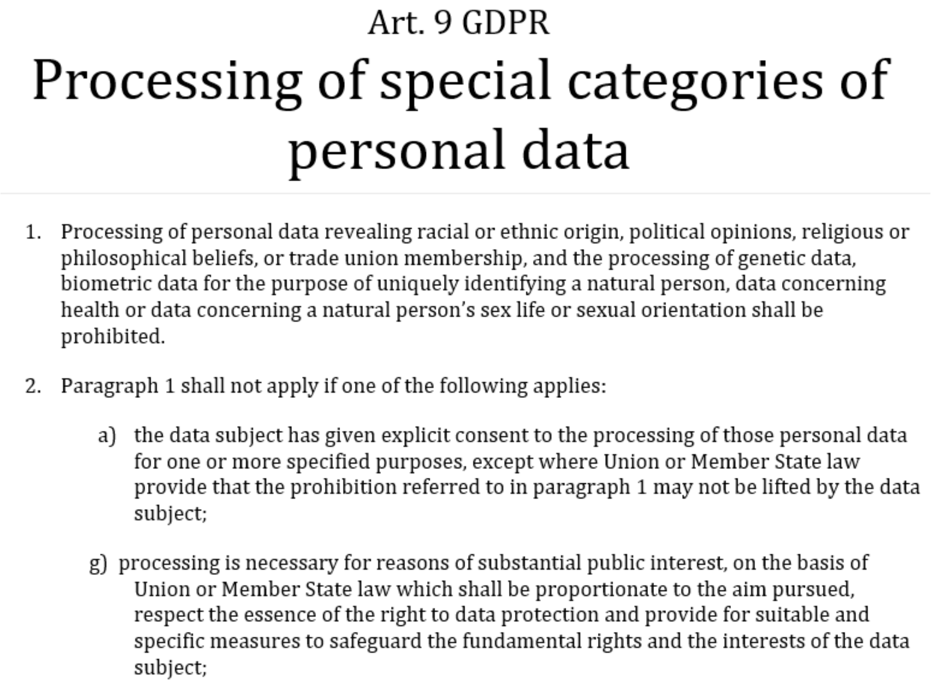
   b. is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

   c. is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Figure 2. Art. 22 GDPR



Art. 9 GDPR

Processing of special categories of personal data

1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

2. Paragraph 1 shall not apply if one of the following applies:

   a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject;

   g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject;

Figure 3. Art. 9 GDPR

1.4.2 Deontic Logics

Deontic logics concern philosophical and formalized logics dealing with obligation, permission, and prohibition. In our context, this means that their normative aspect is home in Ethics, which is theoretically the basis of laws. As we were trying to regulate AI decisions in a post-proof that returns recommended actions, we defined basic rules of what should be realised by a deontic concept of ought in the regard of AI in employment. For this aim, we needed logics that deal with norms and normative systems. Deontic logic is the logic of norm-propositions and it is asking for truth-values[17]. In short, we have the following axioms:

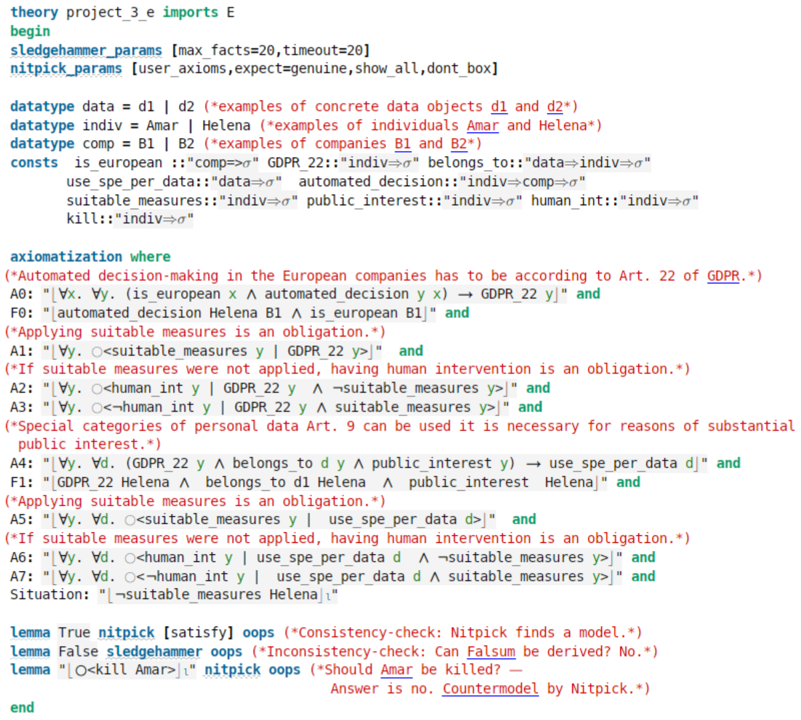
- OB obligatory that,
- PE permissible that,
- IM impermissible/prohibited that.[17]

There are several options for deontic logics. In our case, we are interested in Standard Deontic Logic (SDL), and Dyadic Deontic Logic (DDL)[18]. SDL uses possible world models with serial accessibility relations. SDL is easy and sufficient if the problem can be fully expressively formalized in a set of axioms. However, if there are a lot of conditions and interdependent links, it could be facing paradoxes which can be avoided through DDL. Thus, in the case of violated norms or contrary-to-duty (CTD) issues, additional conditional obligations can be consulted in DDL[3]. We decided to try basically to formalize our problem in SDL, and to integrate E, a DDL developed by Åqvist[19], only if it is necessary to avoid paradoxes or to delineate violated norms.

1.4.3 Implementation

Using the interactive user interface of Isabelle/HOL and in the context of DDL based on first order logic, we defined the obligations and the actions that should be taken if these obligations get violated as following:

$A_1 : \forall y \bigcirc (suitable\_measures\ \ y/GDPR\_22\ \ y)$
$A_2 : \forall y \bigcirc (human\_int\ \ y/GDPR\_22\ \ y \wedge \neg suitable\_measures\ \ y)$
$A_3 : \forall y \bigcirc (\neg human\_int\ \ y/GDPR\_22\ \ y \wedge suitable\_measures\ \ y)$
$A_5 : \forall y \bigcirc (suitable\_measures\ \ y/use\_spe\_per\_data\ \ y)$
$A_6 : \forall y \bigcirc (human\_inty/suitable\_measures\ \ y \wedge \neg suitable\_measures\ \ y)$
$A_7 : \forall y \bigcirc (\neg human\_int\ \ y/use\_spe\_per\_data\ \ y \wedge suitable\_measures\ \ y)$

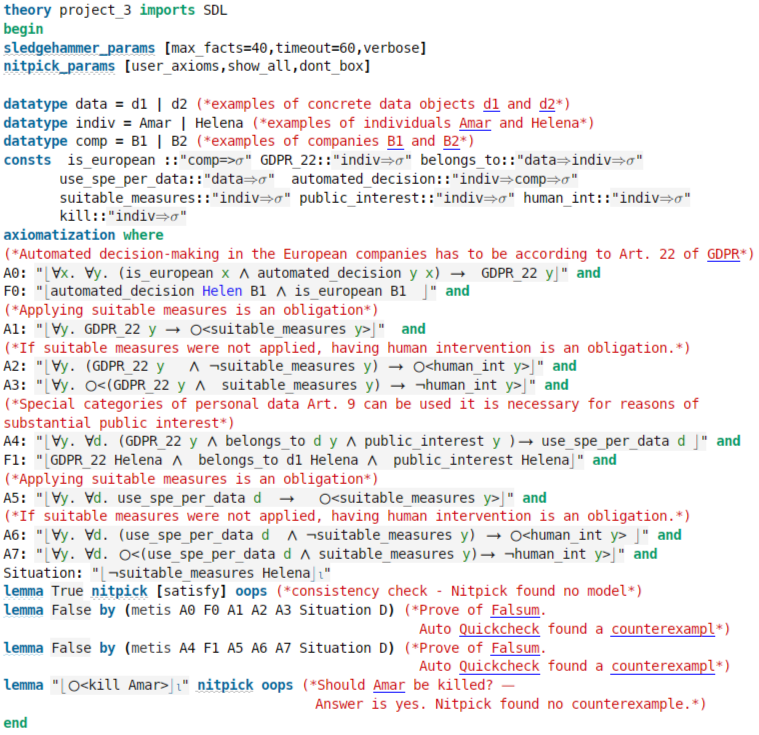


```
theory project_3_e imports E
begin
sledgehammer_params [max_facts=20,timeout=20]
nitpick_params [user_axioms,expect=genuine,show_all,dont_box]

datatype data = d1 | d2 (*examples of concrete data objects d1 and d2*)
datatype indiv = Amar | Helena (*examples of individuals Amar and Helena*)
datatype comp = B1 | B2 (*examples of companies B1 and B2*)
consts  is_european ::"comp=>σ" GDPR_22::"indiv⇒σ" belongs_to::"data⇒indiv⇒σ"
        use_spe_per_data::"data⇒σ"  automated_decision::"indiv⇒comp⇒σ"
        suitable_measures::"indiv⇒σ" public_interest::"indiv⇒σ" human_int::"indiv⇒σ"
        kill::"indiv⇒σ"

axiomatization where
(*Automated decision-making in the European companies has to be according to Art. 22 of GDPR.*)
 A0: "⌊∀x. ∀y. (is_european x ∧ automated_decision y x) ⟶ GDPR_22 y⌋" and
 F0: "⌊automated_decision Helena B1 ∧ is_european B1⌋" and
(*Applying suitable measures is an obligation.*)
 A1: "⌊∀y. ○<suitable_measures y | GDPR_22 y>⌋"  and
(*If suitable measures were not applied, having human intervention is an obligation.*)
 A2: "⌊∀y. ○<human_int y | GDPR_22 y  ∧ ¬suitable_measures y>⌋" and
 A3: "⌊∀y. ○<¬human_int y | GDPR_22 y ∧ suitable_measures y>⌋" and
(*Special categories of personal data Art. 9 can be used it is necessary for reasons of substantial
  public interest.*)
 A4: "⌊∀y. ∀d. (GDPR_22 y ∧ belongs_to d y ∧ public_interest y) ⟶ use_spe_per_data d⌋" and
 F1: "⌊GDPR_22 Helena ∧  belongs_to d1 Helena  ∧  public_interest  Helena⌋" and
(*Applying suitable measures is an obligation.*)
 A5: "⌊∀y. ∀d. ○<suitable_measures y |  use_spe_per_data d>⌋"  and
(*If suitable measures were not applied, having human intervention is an obligation.*)
 A6: "⌊∀y. ∀d. ○<human_int y | use_spe_per_data d  ∧ ¬suitable_measures y>⌋" and
 A7: "⌊∀y. ∀d. ○<¬human_int y |  use_spe_per_data d ∧ suitable_measures y>⌋" and
 Situation: "⌊¬suitable_measures Helena⌋ι"

lemma True nitpick [satisfy] oops (*Consistency-check: Nitpick finds a model.*)
lemma False sledgehammer oops (*Inconsistency-check: Can Falsum be derived? No.*)
lemma "⌊○<kill Amar>⌋ι" nitpick oops (*Should Amar be killed? —
                                        Answer is no. Countermodel by Nitpick.*)
end
```

Figure 4. CDL - E system

In our case, applying suitable measures is obligatory to ensure the rights of the applicant y including processing the personal data lawfully, if it is not the case the program will ask for human interventions. Figure 4 shows the implementation of system E. It is worthy to be mentioned that LogiKEy facilitated the implementation task by enabling importing E a file which provides a Shallow Semantic Embedding (SSE) of a quantified extension of Åqvist's System E in HOL[15].

Nitpick which is an open-source counterexample generator for Isabelle/HOL was used to assess the consistency, and it returned that the identified knowledge base has a model. The command sledgehammer was used to check if falsum can be derived, and it is not the case here.



```
theory project_3 imports SDL
begin
sledgehammer_params [max_facts=40,timeout=60,verbose]
nitpick_params [user_axioms,show_all,dont_box]

datatype data = d1 | d2 (*examples of concrete data objects d1 and d2*)
datatype indiv = Amar | Helena (*examples of individuals Amar and Helena*)
datatype comp = B1 | B2 (*examples of companies B1 and B2*)
consts  is_european ::"comp=>σ" GDPR_22::"indiv⇒σ" belongs_to::"data⇒indiv⇒σ"
       use_spe_per_data::"data⇒σ"  automated_decision::"indiv⇒comp⇒σ"
       suitable_measures::"indiv⇒σ" public_interest::"indiv⇒σ" human_int::"indiv⇒σ"
       kill::"indiv⇒σ"
axiomatization where
(*Automated decision-making in the European companies has to be according to Art. 22 of GDPR*)
A0: "⌊∀x. ∀y. (is_european x ∧ automated_decision y x) ⟶  GDPR_22 y⌋" and
F0: "⌊automated_decision Helen B1 ∧ is_european B1  ⌋" and
(*Applying suitable measures is an obligation*)
A1: "⌊∀y. GDPR_22 y ⟶ ○<suitable_measures y>⌋"  and
(*If suitable measures were not applied, having human intervention is an obligation.*)
A2: "⌊∀y. (GDPR_22 y   ∧ ¬suitable_measures y) ⟶ ○<human_int y>⌋" and
A3: "⌊∀y. ○<(GDPR_22 y ∧  suitable_measures y) ⟶ ¬human_int y>⌋" and
(*Special categories of personal data Art. 9 can be used it is necessary for reasons of
substantial public interest*)
A4: "⌊∀y. ∀d. (GDPR_22 y ∧ belongs_to d y ∧ public_interest y )⟶ use_spe_per_data d ⌋" and
F1: "⌊GDPR_22 Helena ∧  belongs_to d1 Helena ∧  public_interest Helena⌋" and
(*Applying suitable measures is an obligation*)
A5: "⌊∀y. ∀d. use_spe_per_data d  ⟶   ○<suitable_measures y>⌋" and
(*If suitable measures were not applied, having human intervention is an obligation.*)
A6: "⌊∀y. ∀d. (use_spe_per_data d  ∧ ¬suitable_measures y) ⟶ ○<human_int y> ⌋" and
A7: "⌊∀y. ∀d. ○<(use_spe_per_data d ∧ suitable_measures y)⟶ ¬human_int y>⌋" and
Situation: "⌊¬suitable_measures Helena⌋ι"
lemma True nitpick [satisfy] oops (*consistency check - Nitpick found no model*)
lemma False by (metis A0 F0 A1 A2 A3 Situation D) (*Prove of Falsum.
                                                   Auto Quickcheck found a counterexampl*)
lemma False by (metis A4 F1 A5 A6 A7 Situation D) (*Prove of Falsum.
                                                   Auto Quickcheck found a counterexampl*)
lemma "⌊○<kill Amar>⌋ι" nitpick oops (*Should Amar be killed? —
                                      Answer is yes. Nitpick found no counterexample.*)
end
```

Figure 5. SDL

Figure 5 illustrates the implementation when the inference engine is based on SDL. It is shown that we did not use conditional obligations in defining the axiomatization. The data supplied by LogiKEy enables also importing SDL, a file of SSE-based implementation of SDL. Nitpick revealed that no model can be found and it is a well known problem in the case of SDL[3,20].

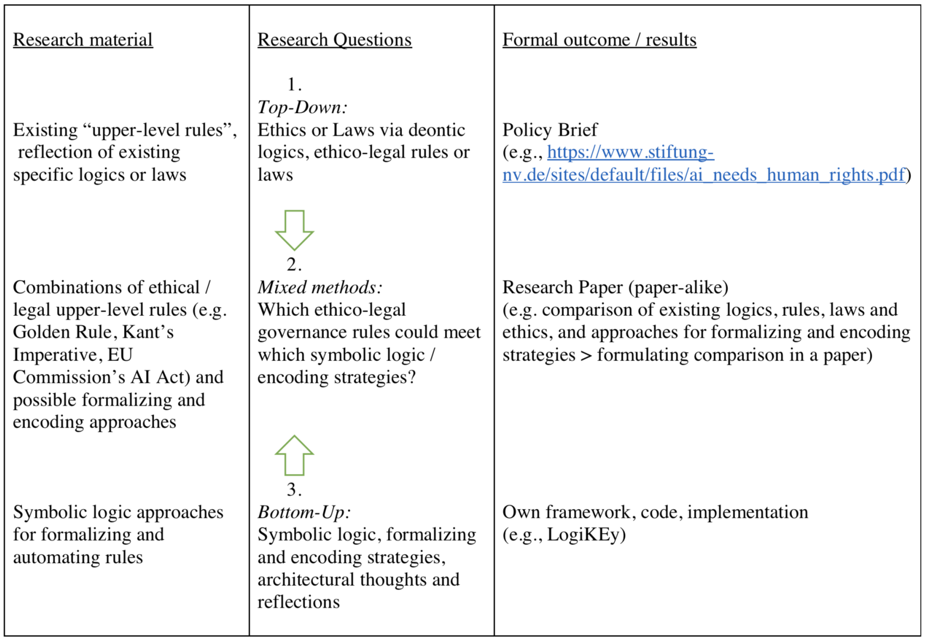
2 Reflection on the interdisciplinary teamwork

In this section, we would like to talk about our experience as a team with different backgrounds during the last three months. We will shortly demonstrate why each of our disciplines is individually relevant for ethico-legal governance in general and refine this relevance in the light of its concrete application in our project.

In our research phase, each of us has familiarized with existing research on the governance of AI based on symbolic AI, summarized and presented it to each other, and discussed it in order to find a specific example for our project to be studied.

It is worth mentioning that the first seminar workshop was a very helpful guide for us not only to refine our ideas, plan the project, and determine our possibilities and limitations but also to think about how we can employ our different backgrounds for working on an interesting project and how to split the tasks accordingly. Because ethico-legal governance can be approached from different levels and topics we needed to choose one specific approach. We did so in roughly conceptualizing the field as depicted in Figure 6.

As we then focused on AI in employment, Data Science and Computer Science students were responsible for providing explanations on some of the decision making algorithms that can be used for hiring an applicant and how they work. On the other hand, for studying the idea of governing of AI systems ethically and legally based on symbolic AI, all of us needed to know more about symbolic AI and its applications. Students of the Humanities made the team familiar with normative ethical theories and focused on studying GDPR as well as the AI Act aiming at determining the possible legal governance of AI in



Research material	Research Questions	Formal outcome / results
Existing “upper-level rules”, reflection of existing specific logics or laws	1. *Top-Down:* Ethics or Laws via deontic logics, ethico-legal rules or laws ⇩	Policy Brief (e.g., https://www.stiftung-nv.de/sites/default/files/ai_needs_human_rights.pdf)
Combinations of ethical / legal upper-level rules (e.g. Golden Rule, Kant’s Imperative, EU Commission’s AI Act) and possible formalizing and encoding approaches	2. *Mixed methods:* Which ethico-legal governance rules could meet which symbolic logic / encoding strategies? ⇧	Research Paper (paper-alike) (e.g. comparison of existing logics, rules, laws and ethics, and approaches for formalizing and encoding strategies > formulating comparison in a paper)
Symbolic logic approaches for formalizing and automating rules	3. *Bottom-Up:* Symbolic logic, formalizing and encoding strategies, architectural thoughts and reflections	Own framework, code, implementation (e.g., LogiKEy)

Figure 6. Initial concept of possible approaches

employment. Their critical additions and questions in our vivid discussions enabled us to choose the level we can broach the project and revealed broader ethical reflections to be considered.

In the embedding phase, the team was split into two interdisciplinary teams:
The first team with backgrounds in Philosophy and Data Science/Mathematics worked on the implementation task using Isabelle/HOL. This cooperation enabled the team to dive deeper in understanding the task: Namely the knowledge in logic, ethics, and mathematical logic played an essential role in comparing different deontic logics, exploring possible solutions for coding the chosen articles, and finally coding them in Isabelle. Complementing the backgrounds facilitated the task, reduced the time needed and the workload in addition to the valuable interesting discussion on both technical and ethico-legal dimensions.

The second team with backgrounds in Computer Science, Data Science and History of Science worked on the proposed example, namely the automatic hiring process, specifying a use case in the workshops conducted in the seminar. Analyzing this example in detail allowed us to better understand all the work done previously on the background of ethical and legal principles for a governance system. From a Computer Science point of view it was interesting to dwell on the application of symbolic logic techniques in real life. Furthermore, it was stimulating to understand what goes beyond simple and abstract programming, and to focus on the ethical part. In its early stage, AI has faced criticism for solely focusing on problem solving and ignoring the socio-technical environment for which it is designed. The historical point of view shows how such criticism is important to be a part of the AI design process aiming at a realistic and effective general-purpose technology.

On the other hand, the goal in the early beginning was to work on a deeper level of governing AI in employment that ensures or at least improves the degree of trustworthiness of the decision making algorithms especially regarding basic human rights of non-discrimination (Article 21 of EU Charter), equality between women and men (Article 23), and the integration of people with disabilities (Article 26)[16]. In the context of this project and in the given time, it was only possible to highlight many challenges to be considered for further work which emphasize the necessity of interdisciplinary work. The chosen articles of GDPR and the approach basing on symbolic AI and deontic logics opened questions on:

- Treatment of personal data: What is in concrete cases considered as sensitive personal data and what is on the other hand necessary for a hiring process?[21]
- How does our use case which is legally binding to GDPR apply in complex cases of companies outside of the EU?[22]
- The need for an explicit definition of suitable measures: Which measures are considered suitable? In which cases should they be applied? How to define the mechanism of governing them?
- Defining actions: How could we define future actions? Action and its execution need to be connected to natural and legal language which is not easy to integrate formally and unified.

- Consistency with other articles of GDPR: How can we reach maximal consistency of GDPR articles? Since laws are often linked to each other and more like a semantic web than a stand-alone article, it is challenging to avoid contradictions and ensure a unified understanding of the legal texts.

Generally speaking, we believe in the necessity of working in interdisciplinary teams in the field of AI, especially for studying forms of governing AI ethically. From our experiences in the previous courses at the university as well as from experiences gained from practice, it is notable that the focus of Data Scientists and Computer Scientists is mostly on finding new algorithms to allow new applications of AI or improving the efficiency of the existing applications - regardless of considering harms that could be caused. However, investigating the (missing) ethics of such algorithms does require knowledge of ethics and Philosophy in general as well as a mathematical understanding of the treated problem in order to match both ethical and technical solutions. We think that this gap could only be bridged by giving ethics of AI a higher priority from the educational perspective of universities, e.g., in providing more courses and opportunities for working on this topic in interdisciplinary teams. Investing more valuable efforts of qualified people in this domain enables a broader discourse to broach several ethical dimensions raised by AI algorithms.

3 Reflection in an ethical context

3.1 Technical reflection

In terms of ethical reflections from a technical perspective, it is necessary to bear in mind that technical entities and properties should mirror ethical values as it also has been stated in the Assessment List for Trustworthy Artificial Intelligence (ALTAI)[8]. In the ALTAI list, seven main principles are defined in order to guide the development and use of AI. They are as the following: Human agency and oversight, technical robustness and safety, privacy and data governance, transparency, furthermore diversity, non-discrimination and fairness, as well as societal and environmental well-being, and, finally, accountability.

Developing an ethico-legal governor shall take all seven principles into account. Especially relevant for the aim of our project, however, is firstly transparency which is including explainability: An ethico-legal governor shall be oriented to embody a 'reason of reason', meaning, a transparent reasoning mechanism that reflects upon automated decisions which are based on non-transparent AI, e.g. ML, and explains them.

Human agency and oversight, secondly, is fundamental to our research question which concerns AI in employment. Not only should human oversight be part of every hiring process, but also is human intervention a possible recommended action of our ethico-legal governor. The decision has to be evaluated by humans generally because it includes the processing of personal and therefore sensitive data. Because the decisions are regarding high risk- actions or -processes, in special cases not only human evaluation but also intervention is needed in order to prevent misuse of data and discrimination of applicants. Therefore, in our project, human agency and oversight are, thirdly, closely related to privacy and data governance.

Diversity, non-discrimination, and fairness are, fourthly, central principles for AI in employment. An automated decision shall not be allowed to be discriminatory: Recommended, governed actions shall base the decision on the ethical values of diversity and fairness. In our project, all ethical principles are referring to this principle as non-discrimination in employment shall be ensured through them, e.g., through transparency.

Fifthly, accountability is important for consolidating ethico-legal governors in its still young state of development as a future central part of AI systems. In the broader perspective, industrial standards could reflect upon accountability in the norming of ethico-legal governors. With this as a basis, standard based agreements between companies, branches, and even production countries would assure some degree of ethical consensus in the development and use of AI.

The two remaining requirements of ALTAI, societal and environmental well-being as well as technical robustness and safety, shall not be forgotten or be prioritized any lower. However, they play a derived role in our concrete case: Societal and environmental well-being are consequences of treating applicants in a non-discriminatory and fair way while technical robustness and safety primarily have to be considered in the development of the AI system itself as they are not exclusive to ethico-legal governors. As an ethico-legal governor is based on symbolic AI, we suppose ethico-legal governors are a measure of gaining interior safety. We furthermore do not expect them to need a lot of resources and therefore also not to mean additional harm to the environment. One could surely reflect upon these two remaining principles more intensely, however, due to the limited scope of our project report we decided to only touch on them.

In total, only in including those ethical values stated in ALTAI which have to be treated as laws by the ethico-legal governor, the first step toward trustworthy autonomous systems can be made. The reasoning capability of symbolic AI which is automated in the ethico-legal governor enables a transparent and explainable approach because actions will be controlled based on ethico-legal constraints meeting regulatory standards.

3.2 Social reflection

As we have seen, the technical reflections have a social basis that is constitutive of future regulatory standards. But in addition to that, further social reflections need to be done.

Firstly, a wider socio-technical context has to be considered: Developers should be aware of the social implications and the according to the context of deployment. It is now a reached consensus that ethical reflections concerning the socio-technical implications of the development and use of AI are still too shallow and theoretical. It seems therefore to be a reasonable claim to work in strict compliance with basic human-centric norms and values which ideally are already manifested in laws[23].

Secondly, it is thus necessary to inform the user or in our case applicant of the decision based on automated systems. As stated before, human intervention should be basic regarding controlling and evaluating actions. This is already being taken into account by the EC and therefore covered by its AI Act[1].

When an individual is subject to an automated decision, however, not only information about it is necessary but also resources to stand up against this decision: Contestability, thirdly, refers to the user's right and ability to argue against an automatically generated decision. It should not be regarded as sufficient to have abstract abilities for doing so in a legal way. A legal approach to defending oneself against an automated decision seems to be too complex and therefore expensive for a citizen as he or she needs to familiarize not only with the legal language but also with the legal landscape. Information and expert support in this regard should be sufficient to enable the person to exercise its right[24].

At last, for now, the reflection on data protection shall be continuous: Is it sufficiently covered by the law or does it need to adapt to changes coming with new developments? How can we prevent the flow of what is defined as sensitive data? Can it be accomplished to reduce the sensitive data to a minimum?[21] Many questions are left to answer.

3.3 Research reflection

Because ethico-legal governance as a research field is very new, future requirements concerning research and development (R&D) shall be included in reflection.

To work interdisciplinary is highly important as there are multiple disciplines involved: Including, but not limited to Jurists who are needed as experts of the legal landscape and its challenges, Logicians and Mathematicians who are needed to develop appropriate logical approaches to complex systems of values and norms whereas, to that end, Ethicists and general Philosophers are needed for broadly reflected pluralistic approaches and definitions, and Computer Scientists as well as Data Scientists are needed to work on methods for automating solutions and increasing compatibility and interoperability with AI systems. Only by including scientists of all relevant fields can the complex problem be approached appropriately.

Furthermore, broader ethico-legal research projects should be initiated as there is a lot of basic research to be done in the next few years to approach the issue properly. An ethico-legal governor should for example be adapted to the capacity of Natural Language Processing (NLP) and Natural Legal Language Processing (NLLP). However, research herein still seems to be in its infancy[25]. Also, the existing basic research on deontic logics, theorem provers, compatibility, and generalizability needs to be expanded. Broader ethico-legal research projects should include different approaches to further ethics and articles in order to reach interoperable ontologies of ethics and laws as a mirror of pluralistic societies.

Conclusion

In this project, we detected the possibility of governing AI systems using symbolic AI techniques where ethico-legal governance refers to processes and procedures that are useful for achieving multiple objectives which meet high performance standards and at the same time evaluate their decisions on the basis of ethical values and laws.

With a risk based focus on using AI in employment where high risk is determined by the potential and probability of harm to an individual, we codified the related articles of GDPR. Using first-order logic and deontic logics we defined obligations and actions to be taken in the event of a breach with those obligations and, with help of LogiKEy, implemented them in Isabelle/HOL by the embedding of ethico-legal rules regarding AI in hiring.

In the light of the seven principles of ALTAI which shall guide the development and use of AI as well as ensure that technical decisions reflect upon ethical values, the technical reflections of our project were discussed. Furthermore, the social and research reflections were broached to introduce and clarify the different ethical dimensions of this work.

References

1. European Commission. Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Proposal for a Regulation of the European Parliament and the Council. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (2021).
2. EU Parliament and Council. General Data Protection Regulation. https://gdpr-info.eu (2016).
3. Benzmüller, C., Parent, X. & van der Torre, L. Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support. *Artif. Intell.* **287**, 103348 (2020).
4. Arkin, R. C., Ulam, P. D. & Duncan, B. An ethical governor for constraining lethal action in an autonomous system. Tech. Rep., Georgia Institute of Technology (2009).
5. Theodorou, A. & Dignum, V. Towards ethical and socio-legal governance in ai. *Nat. Mach. Intell.* **2**, 10–12 (2020).
6. Cath, C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges (2018).
7. Benzmüller, C. & Lomfeld, B. Träumen vernünftige Maschinen von Gründen? Eine reale Utopie. https://www.bbaw.de/files-bbaw/user_upload/publikationen/BBAW_Verantwortung-KI-3-2020_PDF-A-1b.pdf (2020).
8. HLEG on AI & European Commission. The assessment list for trustworthy artificial intelligence (altai) for self assessment. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (2020).
9. Ilkou, E. & Koutraki, M. Symbolic Vs Sub-symbolic AI Methods: Friends or Enemies? DOI: 10.1145/3340531.3414072 (2020).
10. Yalçın, O. G. Symbolic vs. Subsymbolic AI Paradigms for AI Explainability. *Towards Data Sci.* (2021).
11. Dickson, B. What is Symbolic Artificial Intelligence? *TechTalks* (2019).
12. Benzmüller, C. & Lomfeld, B. Reasonable Machines: A Research Manifesto. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, 251–258 (Springer, 2020).
13. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (2018).
14. Mozur, P. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html (2019).
15. Benzmüller, C. *et al.* LogiKEy workbench: Deontic logics, logic combinations and expressive ethical and legal reasoning (Isabelle/HOL dataset). *Data brief* **33**, 106409 (2020).
16. European Union. Charter of Fundamental Rights of the EU. https://www.europarl.europa.eu/charter/pdf/text_en.pdf (2000).
17. McNamara, P. & Van De Putte, F. Standard Deontic Logics. https://plato.stanford.edu/entries/logic-deontic/ (2006).
18. Benzmüller, C., Farjami, A. & Parent, X. Faithful Semantical Embedding of a Dyadic Deontic Logic in HOL. https://arxiv.org/abs/1802.08454 (2018).
19. Benzmüller, C., Farjami, A. & Parent, X. Åqvist's dyadic deontic logic E in HOL. *J. Appl. Logics–IfCoLoG J. Logics their Appl. (Special Issue on Reason. for Leg. AI)* **6**, 733–755 (2019).
20. Parent, X., van der Torre, L. & Sun, X. Handout Lecture 1: Standard Deontic Logic. *unpublished notes* (2016).
21. European Commission. What personal data is considered sensitive? https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en (2016).
22. EU Parliament and Council. Does the GDPR apply to companies outside the EU? https://gdpr.eu/companies-outside-of-europe/ (2016).
23. Lorenz, P. & Saslow, K. Artificial Intelligence Needs Human Rights. https://www.stiftung-nv.de/sites/default/files/ai_needs_human_rights.pdf (2019).
24. Lyons, H., Velloso, E. & Miller, T. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proc. ACM on Human-Computer Interact.* **5**, 1–25 (2021).
25. Robaldo, L., Villata, S., Wyner, A. & Grabmair, M. Introduction for artificial intelligence and law: special issue "natural language processing for legal texts" (2019).