LEGALIS
Artificial Intelligence Instruments to Assist Regulators in Assessing the Legal Compliance and Ethical Alignment of Artificial Intelligence Systems

Research Abstract

The rapid advancement of artificial intelligence (AI) in various sectors has brought significant benefits but also raised concerns about risks and ethical considerations. Initiatives like the EU AI Act and the U.S. government's executive orders highlight the importance of ensuring that AI systems are lawful, ethical, and technically robust. In response, the LEGALIS research program aims to develop fundamental AI tools to assist regulators in evaluating the legal compliance and ethical alignment of AI systems before deployment. This endeavour calls for challenging advances of the state of the art in multiple AI fronts to produce the technical means: (i) to represent legal and ethical knowledge; (ii) to learn a model (of the behaviour) of an AI under inspection; (iii) to compute legal and ethical compliance of the inspected AI; and (iv) to explain to stakeholders the result of the compliance analysis as well as of the behaviour of an inspected AI.

These objectives aim to equip regulators with essential tools for assessing the legality and ethics of AI systems. The technological objectives include building a demonstrator to showcase project results and validating them with legal and ethics experts.

The LEGALIS research program addresses critical challenges in verifying the legality and ethics of AI systems, aiming to foster trust and safety in AI deployment. Through collaboration with experts and empirical evaluation in two real-world domains (automated driving systems and social assistive robotics), LEGALIS strives to advance the field of trustworthy AI and contribute to responsible AI development and deployment.

A/ Research background and objectives

Research Programme

Motivation

Artificial intelligence (AI) has achieved notable momentum in the last decade, enabling impressive results. Indeed, the deployment of AI is expanding in our lives in a broad range of industries, such as education, healthcare, finance, logistics, mobility, and law enforcement [Vin+20]. However, alongside such growth, the risks posed by AI have become apparent, from using it to fool systems with malicious purposes or produce cyber-physical attacks, to name but a few. Consequently, there has been an increasing interest in handling such threats by promoting a trustworthy, human-centred AI [Cha21]. Along these lines, the European Union has established its vision of trustworthy AI [EC23], which should be: (1) **lawful**, respecting all applicable laws and regulations; (2) **ethical**, respecting ethical principles and values; and (3) **technically robust and safe**. This concern for trustworthy AI goes beyond the European borders. In January 2023, the National Institute of Technology and Standards (NIST) published its Artificial Intelligence Risk Management Framework (AI RMF 1.0). The Framework identifies the core characteristics of trustworthy AI, equips AI actors (organisations and individuals) with approaches that increase the trustworthiness of AI systems and is designed to help foster the responsible design, development, deployment, and use of AI systems over time. This effort has been endorsed by the U.S. government, which has also recognised the imperative to establish guardrails for the responsible development and deployment of AI technologies. In this manner, on October 30, 2023, President Biden issued an executive order to ensure that the U.S. manages the risks of artificial intelligence, emphasising references to generative AI [Fac23]. These measures are meant to ensure AI systems are safe, secure, and trustworthy before companies make them public and even during their deployment. The NIST plays a central role in that executive order since it is tasked with developing standards for trustworthy AI. More recently, the European Parliament took a step further towards AI regulation, which goes beyond standards: AI in the EU will be regulated by the AI Act, the world's first comprehensive AI law [EC24].

On the research side, the threats posed by AI have spurred a wealth of research on trustworthy AI. Indeed, there has been much research in AI & safety and security (e.g., [Amo+16, Lei+17], [Cam+24]), and, more recently, AI & ethics have also attracted considerable attention (e.g., [Dig+18, Dig19]), whereas the lawful dimension has yet to be explored. Furthermore, as transparency is also an intrinsic condition for trust and trustworthy AI, AI research has tackled how

AI systems should explain their decisions in a manner adapted to the concerned stakeholders [Mil19]. This endeavour has been particularly necessary with the advent of very successful deep learning models, including LLMs, which are notoriously hard to explain [Ras+23].

Against this background, the recent advances in regulation, exemplified by the acts mentioned above, suggest that regulators entered the game. At this point, a major question driving our LEGALIS proposal arises: ***how to verify that a given AI system is lawful and ethical?*** This is a rather intricate issue as it requires regulators to be capable of thoroughly inspecting an AI system to verify whether its behaviour abides by the applicable laws and ethical principles. Although, as prescribed, e.g., in the White House's executive order [Fac23], AI developers must share their safety test results and other critical information with the government, we argue that it is still the government's responsibility to verify the trustworthiness of AI systems before and after an AI system deployment. This proposal will pursue the provision of relevant concepts, methods and tools to support regulators in such tasks. Furthermore, we argue that AI methods can be useful to assess the trustworthiness of AI systems. In other words, through LEGALIS we aim to exploit AI to equip regulators with the technological means to verify, approve and, possibly, improve AI systems before they are deployed in the real world.

To the best of our knowledge, there are no AI instruments that assist legal experts and ethics experts in assessing the legal and ethical compliance of AI systems alike. Therefore, this project aims at filling this gap. We are aware of the development of corporate and business compliance in the last twenty years, after the introduction of the Sarbanes-Oxley Act (SOX) in 2002 to improve auditing and public disclosure in response to numerous accounting scandals in the early 2000s (e.g., Enron or Arthur Andersen) (see https://www.law.cornell.edu/wex/sarbanes-oxley_act). This fostered the Governance, Risk management and Compliance (GRC) field to pursue integrity, i.e., that organisations reliably achieve objectives aligned with ethics and law. We can learn and draw upon from this experience [Has18], but our scope is much broader: our aim is not corporate auditing but anticipating and testing AI systems, making sure that they comply with the ethical and legal requirements that make them trustworthy and safe.

Global objective

As a consequence, LEGALIS ***global objective*** is

To develop the fundamental AI instruments that assist regulators in assessing the legal compliance and ethical alignment of AI systems before their large-scale deployment in the real world.

Detailed objectives

Although, as required by the FBBVA PRISMS&PROBLEMS call, the motivation of this proposal is practical, the achievement of our global objective above poses novel foundational research challenges that the LEGALIS team will address. Hereafter, we will differentiate between two types of objectives in which we decompose our global objective: research (RO) and technological (TO) objectives. Figure 1 graphically outlines a process-oriented view of the proposed LEGALIS framework. It will also help us explain the dependencies between research and technical objectives which revolve around developing theoretical and algorithmic tools for analysing and assessing the legal and ethical dimensions of an AI system. Furthermore, since we propose to exploit AI for such purposes, we also consider the means for our AI tools to guarantee the "right to an explanation" that the EU AI Act imposes on AI systems. We will turn them "explicable", i.e., "explainable" and "accountable" alike [Flor19].

For generality, and to focus our research and technological objectives, we propose to consider two actual-world case studies: **automated driving systems** (ADS) and **social assistive robotics**. On the one hand, partner ICMAT-CSIC counts on demonstrated expertise in research on ADS ([Cab21, Ins22, Nav22, Cab23, TRUSTONOMY[1]]). On the other hand, partner IRI-CSIC leads research on social assistive robotics (Can21, ROB-IN[2], TRAIL[3], FRAILWATCH[4]).

On the one hand, we shall focus on supporting automated lane-changing decisions in heterogeneous traffic, which is, and will be, one of the most challenging problems in relation to ADS. Furthermore, it reflects the interplay between legal (with, among others, the EU AI Act and various related ISO and SAE standards) and ethicall (with, among others, the safety vs performance tradeoff and the fact of potentially putting at risk the lives of passengers and pedestrians) that motivate LEGALIS.

On the other hand, we will consider the problem of deploying an autonomous social assistive robot in elderly homes to perform fundamental assistive tasks, such as delivering objects to patients upon request or issuing personalised reminders. Additionally, the robot continuously collects data throughout the day observing the patient to enhance the overall care experience. There are a large number of candidate ethical issues in this scenario (e.g., disclosure of information, degree of influence of the caregiver directives in the decision-making, robot limiting the elder's autonomy, etc). There are also multiple legal issues (normative requirements) to consider in this scenario (e.g., safety, user protection, liability) coming from different sources: EU and national laws, case-based law (judicial decisions), national and EU health and disability policies, and specific constraints set by Hospital ethics committees.

[1] TRUSTONOMY: Build acceptance and trust in autonomous mobility
[2] ROB-IN: Robot for continual personalized assistance able to explain itself. (PLEC2021-007859)
[3] TRAIL: TRAnsparent InterpretabLe robots. (HORIZON-MSCA-DN-101072488)
[4] FRAILWATCH: Frailty robotic monitorization system for elder population (23S06141-001)

Within this background, we propose to address the following objectives.

(RO1) Representing legal and ethical knowledge. Given an AI to inspect a particular domain, our first objective must be to determine the laws that the AI is expected to comply with. For that, a regulator must consider a vast number of enforceable *hard law* provisions (e.g., Parliament Acts and case-based law from the Judiciary, for instance the EU AI Act), policies (e.g., requirements from Government agencies) and *soft laws* (e.g., standards like the NIST AI RMF) [Cas24a,b,c]. They are available as documents contained in large legal repositories (such as B.O.E, Legifrance or Eurlex at the EU level).

Given the large amount of legal documents to handle, we propose to exploit the Large Language Models (LLMs) technology. More precisely, we propose to investigate the use of Retrieval-Augmented Generation (RAG) for LLMs since RAGs have arguably emerged as the only (known) solution to incorporate external knowledge into LLMs to handle common problems such as hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes [Gao+24]. The main research objective is twofold. First, we will explore the available RAG architectures and evaluate/validate them with the help of legal experts to find the most accurate one for retrieving **relevant**, applicable laws. By relevant we mean those laws (out of all applicable laws) that legal experts deem as worth auditing. Therefore, an important aspect to address will be how to incorporate legal experts' feedback (human feedback). This leads to our second objective: how to explore candidate techniques for fine tuning an LLM with legal experts' feedback (Reinforcement Learning with Human Feedback [Chr17], and the more recent and promising Direct Preference Optimisation [Raf24]).

After relevant laws are retrieved for a given domain, we propose to encode them as *legal rules* in some formal language (e.g., deontic-event calculus [Has22, Cas24a]), so that they can be subsequently verified. Therefore, an additional research objective will be to find an appropriate method for translating applicable laws into (legal) rules. This objective has been already faced from a symbolic AI approach, non-standard deontic logic, semantics, LegalRuleML and the *Regorous* methodology [Has15, Gov15, Gov16, Gov17, Lam19]. Thus, we will resort to these tools.

As to representing ethical knowledge, this is an engineering exercise that starts from some *taxonomy of ethical objectives* (e.g., the taxonomy for automated driving systems in [Cab21], or the taxonomy for socially assistive robotics in [Par21]). From that, applied ethics experts will collaborate with AI experts to define ethical values and the corresponding preference model. Here we take the stance that ethical values are moral principles that assess the goodness of actions. For instance, safety can be defined as a value that evaluates negatively the action of running over pedestrians and positively the action of not getting too close. Then, they can proceed to determine the preferences over ethical values (e.g., safety is preferred to comfort) to specify a *value system* [Cab21, Ser23, Mon22]. Ultimately, the AI to inspect is expected to behave

in alignment with that value system, and thus, the resulting value system needs to be tailored for the domain at hand.

(RO2) Learning a model of an AI to inspect. When there is no direct access to the model of the AI under inspection, we can still learn a model of its behaviour by observation. Although there is a large number of learning models from observation in the AI literature, it is important to guarantee model explainability to foster social acceptance and reliability [Mar09]. Based on recent, promising work in [Mel+21, Mel+24], our objective will be to explore how to use the framework of Inductive Logic Programming (ILP) [Mug91] to learn the rules that describe the behaviour of an AI under inspection. Preliminary work of the research team [Ver+23] indicates that ILP is valuable to learn an AI model from observation (of an autonomous car). The resulting rules will be valuable to generate concise and interpretable explanations.

(RO3) Computing legal and ethical compliance. We differentiate two sub-objectives: (i) verifying whether an AI abides by applicable laws, encoded as legal rules (**legal compliance**); and (ii) verifying whether an AI behaves in alignment with a value system as specified by an applied ethics expert (**ethical compliance**).

To check legal compliance, we can start from the result of learning the behaviour of an AI from observation (RO2), or instead consider whether an AI model is already available for inspection (typically as a deep neural network ,DNN). For example, consider our cases above: the AI model of a car might not be made available by the car maker for inspection, whereas the robot maker might design the DNN controlling the robot available for inspection.

In the first case, an AI model needs to be learned from observations (RO2). Assuming that we learn a convenient explainable model of the AI under inspection, based on logical rules, our goal to check legal compliance would amount to verifying whether the learned rules comply with legal specifications from experts (e.g., via model checking [Rob23]). The second case calls for a different approach because a well known and critical issue for DNNs is the lack of robustness to very small variations of the input that may result in significant variations of the output, sometimes resulting in unexpected or undesired decisions. This phenomenon is known as adversarial input [Sze+13, Ami+23, Ins+23]. Considering this, our objective will be to investigate how to exploit adversarial machine learning to generate inputs that help detect non-compliance of legal rules. In a related fashion, given detected non-compliances we could explore minimal changes leading to actual compliance in line with algorithmic recourse developments [Kar+22] or more classical results in decision theoretic sensitivity analysis [Ins90].

Regarding ethical compliance, our objective will be to exploit a specification of ethical values and value systems to compute the degree of *value alignment (ethical compliance)* [Gab20, Mon22, Bro21] of an AI under inspection. Furthermore, an additional goal will be to compute the *behaviour* of an AI with optimal value alignment. Thus, we can give regulators an ideal reference behaviour to help them understand how an ethical AI should behave. For that, we can leverage recent results by the LEGALIS team, which allow us to train an AI to learn how to behave with

optimal value alignment [Nor23a, Nor23b, Rod20, Rod21, Rod22, Rod23] and to learn an explainable rule-based AI model (RO2) for checking ethical alignment similarly to legal compliance [Ver+23].

As a preliminary screening step, we shall also assess the system from a risk management perspective adapting the methods proposed by the research team in [CAM24] for systems including AI-based components.

(RO4) Explaining an AI and its legal and ethical compliance to relevant stakeholders. To meet the stakeholders' right to an explanation, we will pursue two sub-objectives: (i) building *counterfactual* explanations that help understand the behaviour of the AI under inspection; and (ii) explaining the results of assessing legal and ethical compliance.

A counterfactual explanation [Mil19] describes a causal situation in the form: "If X had not occurred, Y would not have occurred". For example: "If I hadn't sipped this hot coffee, I wouldn't have burned my tongue". Event Y is that I burned my tongue; cause X is that I had a hot coffee. Thinking in counterfactuals requires imagining a hypothetical reality that contradicts the observed facts (for example, a world where I have not drunk the hot coffee), hence the name "counterfactual". In interpretable AI, counterfactual explanations can be used to explain decisions in AI tasks. A counterfactual explanation of an AI's decision describes the required changes the AI should consider to reach an alternative decision.

Psychological literature shows that humans understand the world using causal models [Slo05], which formulate the dependencies among multiple factors [Pea+16]. Recent studies have started to use causality to produce explanations for AI systems [Mad+20, Yue+23]. Thus, our goal will be to exploit causal reasoning to yield counterfactual explanations that help stakeholders understand the behaviour of the AI under inspection. Some initial ideas within the LEGALIS team may be seen in [Mei+24, Lov+24]

Our second goal will be to explain the results of the legal and ethical compliance assessment. Since we propose to use interpretable techniques to assess compliance, this goal becomes challenging from a human-computer interaction perspective, since we will aim at explaining the results in a comprehensible manner. We will explore how to build *visual explanations* highlighting non-compliance with laws and misalignment with ethical value. We shall also recover classic work in decision theoretic sensitivity analysis to provide explanations [Ins90].

Addressing the research objectives above will lead us to develop the fundamental AI algorithms required to compute legal and ethical compliance. Furthermore, we will count on the means to explain the behaviour of the AI to inspect and the results of assessing legal and ethical compliance. Furthermore, given the applied nature of this call, and the fact that our AI tools are meant to assist regulators, we propose to demonstrate our results in our two case studies and validate them with the aid of legal and ethical experts, as specified through the following technological objectives.

(TO1) Building a demonstrator that explains legal and ethical compliance. We will build a software application that will demonstrate the technological results of the project. We will equip the demonstrator with different AI models of ADS, each exhibiting different degrees of legal and ethical compliance. Along the same line, we will incorporate different AI models of a social assistive robot. The demonstrator will allow a stakeholder to inquire about the behaviour of a vehicle or a robot and receive counterfactual explanations. It will also allow a user to perform a legal and ethical assessment. The demonstrator will report the results of the assessment.

(TO2) Validating the demonstrator with the aid of experts on law, ethics and responsible AI. With the aid of the legal expert in the project (Prof. Pompeu Casanovas, IIIA-CSIC), the applied ethics expert (Prof. Begoña Román, UB), and the expert on policy making (Dr. Pablo Noriega), we have organised an advisory board with experts on Applied Ethics, Law & Technology and Responsible AI (see work plan section and supporting letters) that will help us validate the results of the project. We will showcase our demonstrator to our advisory board (M18) for validation purposes, with sufficient time (six months) to address the feedback received.

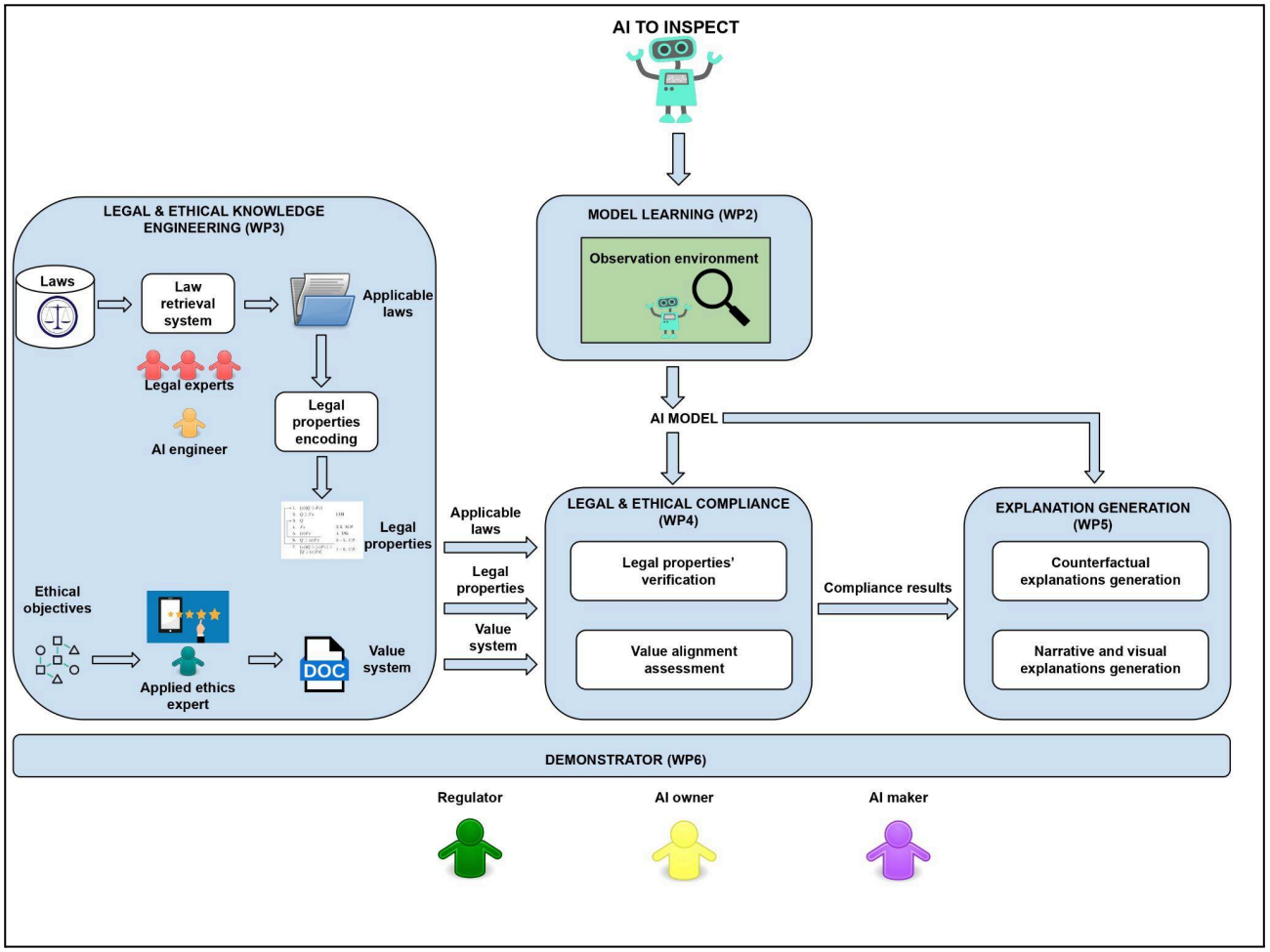

Figure 1. The LEGALIS framework in a nutshell.

Relevance to the call

The proposal LEGALIS fully aligns with the ***FBBVA PRISMS & PROBLEMS*** call within its topic ***Artificial intelligence values, regulation and applications***.

Indeed our team integrates an interdisciplinary group with well-known specialists in Artificial Intelligence, Robotics, Decision Theory, Machine Learning and Statistics, Law and Philosophy (Applied Ethics) that aims to provide a radically new, robust and well-founded AI-based approach and tools to assist regulators in assessing and, if required improving, the legal

compliance and ethical alignment of AI systems before their large-scale deployment in the real world.

Stemming from their extensive multiperspective vision of the state-of-the-art in AI, the LEGALIS team will focus on two massive cutting-edge AI applications, ADS and social assistive robotics that, on the one hand, promise to change the world as we know it, and, on the other, operate in critical domains replacing humans. Safety, security and trustworthiness are key properties in both application domains, constrained as well by the need to comply with an increasingly complicated legal environment (from the EU AI Act, to various cybersecurity norms going through several risk analysis standards) and nuanced ethical issues affecting life and the environment, impacted as well by various cultural lifestyles.

Our innovative tools will serve to facilitate assessing such needs and, very importantly, suggest improvements whenever required. By testing developments within two domains we expect to be able to robustify our tools and generalise to any domain including the currently popular generative AI systems which, incidentally, we shall be using profusely in LEGALIS.

Novelty and Innovation

The fields of ethics and AI and AI and law are relatively young though a number of research lines have started to emerge in the last few years, a number of conferences and journal special issues are now dedicated to this topic, and their origins go back to the eighties and nineties [Ben12] . In what follows, we focus on the most relevant work to the topic of this proposal. From a research perspective, we plan to advance the following research topics.

Value engineering and value alignment. So far, most research works on trustworthy AI [Cha21, Kau22] have signalled general values –such as fairness or transparency– that should be considered when designing any AI system. However, when tailoring to specific domains (e.g., ADS or healthcare), it becomes also necessary to pinpoint those ethical considerations that are specific for the given domain and tailor the value system to fit the specific context. This involves the characterisation of the high-level ethical principles that are relevant in a given domain and its subsequent translation into an operational mathematical formalisation of the moral values that should guide the behaviour of an AI system that is meant to be deployed in such a domain. Then, preferences among such values have to be defined. The literature is very limited in this regard, and thus, as stated in our RO1, we aim at advancing the state of the art of Value Engineering by integrating diverse (interdisciplinary) perspectives from AI experts, decision theorists, ethicists, and domain specialists in task T3.4 (Value System encoding). Subsequently, value alignment aims to ensure that an AI system's behaviour aligns with the engineered value

system. Consequently, it can also assess how closely an AI system adheres to the desired ethical principles that have been formalised for a specific domain. Although value alignment has been extensively discussed in the literature [Gab20, Mon22, Bro21], there is a lack of approaches that operationalise the evaluation of whether the system's actions align with the defined ethical values and its quantification. Some of our initial ideas stemming from a recent proposal in relation to risk management in AI based systems and components may be seen in [Cam+24]. Thus, as RO3 details, we pursue to advance the state of the art by providing methods for the computation of the degree of value alignment (ethical compliance) of a behaviour. In particular, in task T4.2, we plan to evaluate whether the system's actions align with the defined ethical values and we also aim to compute the behaviour of an AI system with optimal value alignment. This serves as an ideal reference for regulators, helping them understand how an ethical AI should behave.

Generative models for legal reasoning. This is a new AI & Law field of research, based on previous work by Bench-Capon, Atkinson and Ashley [Ash17], among others. Legal information retrieval has been mainly developed by publishers (such as Lexis-Nexis, Wolters Kluwer and Westlaw (Thomson-Reuters), the World Legal Information Institutes (e.g., LLI at Cornell), and AI & Law scientific associations, such as the International Association for Artificial Intelligence and Law (IAAIL) and JURIX. Legal reasoning has been largely based so far on argumentation schemas, defeasible deontic logic and legal theory [Rot23]. LegalXML and LegalRuleML have also been developed along with legal ontologies for information retrieval on the semantic web [Cas16] [Oli19].

More recently, researchers have started to investigate the potential benefits of employing LLMs for legal reasoning. Specifically, the Codex (Stanford) team has started exploring the capabilities of LLMs to interpret and apply legal provisions in specific areas, such as tax law. There is a whole research programme to develop it, named "Law informs Code" [Nay22]. In a subsequent work, [Nay24] conducted some guided experiments testing (via different *prompting* strategies) the ability of different LLMs in answering well-defined multiple-choice legal questions sampled from CFR and US Code exams, whose data is equipped with ground-truth answers that can be employed to objectively verify the accuracy of the model. The authors conclude that current models are not ready as a fully *automated replacement of legal experts* (LLMs "as attorneys") in terms of accuracy. Nonetheless, while not considering nor conducting any experiment with "humans-in-the-loop", [Nay24] envision the potential for LLMs to *assist* humans in more complex legal tasks, such as assessing the legal compliance and ethical alignment of AI systems, i.e., the central purpose of this project. The authors also encourage future research on methods that improve LLM legal analysis skills, by designing separate models that can apply legal and ethical standards to confirm whether or not an AI is properly aligned with the law.

In addition to prompt engineering, *fine-tuning* LLMs is often employed to improve the performance of pre-trained models when considering domain-specific tasks. In the context of legal reasoning, to the best of our knowledge the only existing attempt is represented by the unpublished work by [Yue23], who propose the use of classical supervised fine-tuning in the

context of the Chinese legal system to solve objective (similarly to [Nay24]) and subjective question-answering legal tasks. Nonetheless, the authors report that their "LLM [fine-tuned] with high-quality instruction data [...] might produce inaccurate responses due to hallucinations or outdated knowledge". Indeed, as argued by [Gao+24] (see Figure 4 of their article), a fine-tuning approach might not be appropriate in scenarios (like legal reasoning) where accessing a large body of external, up-to-date knowledge is crucial, since it would require additional retraining every time the knowledge base is updated. This is very impractical also considering that classical supervised fine-tuning is very expensive, both in economic terms (e.g., [Yue23] report experiments employing 8 A800 NVIDIA GPUs, for a total cost of ~200.000€ for the training hardware only) and in terms of effort devoted to gathering and curating the large amount of data required by the (supervised) fine-tuning training. For these reasons, [Yue23] also investigate the use of RAG to equip their LLM with an external data source, but they only test a "naive" RAG architecture whose impact is not thoroughly evaluated in a systematic way.

In our project, we also aim to employ RAG to avoid the limitations of fine-tuning, but in contrast with [Yue23], we will systematically investigate which, among the many existing RAG architectures [Gao+24] is more appropriate for the task we tackle, i.e., assisting law experts in the process of determining which laws are applicable to a given AI system (and later verify whether such an AI system comply with those applicable laws). Notice that our task is inherently more complex than legal question-answering as considered in [Nay24] and [Yue23], for which ground-truth data is available to objectively determine the accuracy of the LLM answers. In contrast, in our case the *interpretation of legal experts* is crucial to correctly determine the body of applicable laws. Along these lines, our approach will definitely encompass experts-in-the-loop, to ensure the necessary human oversight and to overcome the drawbacks of the above-mentioned attempts of designing fully-automated legal reasoners.

A well-known approach to incorporate the feedback of humans (legal experts in our case) into LLMs without the drawbacks of classical supervised fine-tuning is by means of reinforcement learning, i.e., Reinforcement Learning with Human Feedback (RLHF) [Chr17]. Nonetheless, as noted by [Raf24], RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning. Along these lines, in this project we will explore the use of the more recent Direct Preference Optimization (DPO) [Raf24], which has been shown to be more stable, performant, and computationally lightweight than RLHF. The advantages of DPO are due to its simplicity, consisting of only two steps: (i) gather a preference dataset with positive and negative selected pairs of generation, given a prompt and (ii) maximise the log-likelihood of the DPO loss directly, hence eliminating the need for fitting a reward model, sampling from the LM during fine-tuning, or performing significant hyperparameter tuning.

Taking advantage of our legal experts, as well the preference models developed in relation to value alignment, we shall also propose fine tuning models through reinforcement learning with human feedback as well as direct preference optimisation, drawing on recent work from the participants [Gall24].

Legal compliance as formal verification plus formal validation. LEGALIS will analyse legal and ethical compliance of AI systems at two separate levels. First, we will exploit techniques for automatic formal verification of complex AI systems, e.g., DNNs [Ami+23]. Then, we will perform formal validation with the help of internal experts (in the research team) and external experts (in an advisory board), in order to assess the effective compliance of the AI system with moral and legal requirements.

As for formal verification, we will first investigate the case where the AI system is represented as a DNN. This type of AI models are completely opaque to human understanding, and they are often sensitive to input perturbations (adversarial input) [Sze+13, Ami+23, Ins+23]. LEGALIS will then extend the state of the art in this field, applying DNN formal verification techniques to the currently unexplored legal and moral context. In particular, as proposed by RO3 (and specifically by task T4.1 in the work plan), we will leverage on recent results attained by our research team [Mar+24] to enumerate input ranges with probabilistic safety guarantees, according to user-defined safety properties (rules). Thanks to the probabilistic relaxation, our approach can scale to the rich semantics of legal rules (defined with the help of internal experts), providing useful hints about the legal fallacies of the AI system.

While adversarial input analysis provides valuable information from the legal perspectives (where obligations are involved), the study of moral compliance requires a different formal verification approach. Indeed, moral rules typically express preferences over a set of possible behaviours, rather than strict safety properties. As a consequence, this problem is often regarded as ranking inputs according to ethical compliance, instead of discerning between allowed and not allowed inputs [Den16]. In this sense, it is more convenient to develop a logical representation of the AI system [Den16], learned, e.g., via Inductive Logic Programming (ILP) [Mug91] as proposed by research objective RO2 (and specified by task T2.2 in the work plan). Formal verification will then be realised performing automated reasoning on the learned logical model, and ranking possible flows of execution of the AI system according to the moral principles defined by stakeholders. To this aim, we will exploit the rich semantics of Answer Set Programming (ASP) [Lif19], which we already used to express multiple moral values for autonomous agency via preference reasoning and optimization (weak) constraints [Ver+23].
The ILP approach is also useful in case the DNN (or equivalent) model of the AI system is not available in advance, and then an explainable comprehensive model shall be derived. In this way, we will then extend the state of the art in verifying moral [Den16] and legal [Rob23] alignment via model checking, jointly verifying the compliance of the ASP model of the AI system to legal rules (expressed as constraints in ASP) and ethical preferences.

As for *legal validation*, we will separate legal compliance from business and corporate compliance [Mcg23]. We will advance the state of the art by incorporating a validation regulatory methodology that sits on top of formal verification. Thus, LEGALIS will also consider the *legal validation* of applicable laws. Validation of legal validity (legality) of regulatory or normative systems (applicable laws) will be undertaken by legal experts (in the research team

and in the advisory board). These will validate the legality of the output produced by the formal verification process by defining thresholds and degrees of compliance [Lam18].

Model learning from observation. Explainability is a crucial requirement for machine learning models supporting decision-making in safety-critical contexts, or whenever validation and trust from human users is of utmost importance [Mar09]. In particular, machine learning methods which are explainable by construction are powerful to generate a human-graspable approximation of the underlying process, mimicking the cognitive flow of humans [Bur21].

In LEGALIS, we will devise methods to learn a rule-based model from observation, relying on Inductive Logic Programming (ILP) [Mug91]. ILP has been successfully adopted in numerous scenarios, including natural language processing to program analysis and robotics [Cro22]. In particular, we will extend the use of ILP to derive an explainable legal model from observations. To this aim, we will rely on a suitable rule formalism for legal knowledge representation, e.g., deontic logic to reason about obligations [Von51]. Deontic theories can be implemented in several logic programming paradigms, out of which Answer Set Programming (ASP) [Lif99] has been shown to be very efficient for compliance checking [Rob23]. However, deontic theories are still hand-written, thus limiting their applicability in real use cases subject to complex normative systems. Hence, we will extend existing tools for ILP under the ASP semantics [Law20] to learn complex deontic operators [Gio13], in order to well represent realistic normative systems. In this direction, we will leverage our expertise in learning ASP programs encoding ethical values [Ver+23].

Explanatory Artificial Intelligence (XAI). In line with our comments above in relation to counterfactuals, we shall explore how such ideas may be used to explain legal compliance and ethical alignment of AIs. Our approach will look for minimal input changes leading to a modification of a compliance recommendation (if positive to suggest the most relevant inputs in determining the recommendation; if negative, to suggest changes to reach positive compliance) or an alignment assessment (to suggest most sensitive change directions leading to improvement or worsening of alignment). Initial ideas within the robotics domain may be seen in [Lov+24]. For this we shall use recent tools from causal inference {Mei+24] and algorithmic recourse [Kar+22] as well as draw on work in decision theoretic sensitivity analysis to explain optimality of decisions [Ins90].

Experimental software tools for analysing the legal and ethical compliance of AI systems. Finally, from a technological perspective, we are aware of the many surveys available for

compliance[5]. Nonetheless, to the best of our knowledge, there are no software tools that currently support the verification of legal and ethical compliance of AI systems. Our purpose in this project is to make headway, from an experiential perspective, along this direction. Ultimately, in WP6, we aim at developing AI instruments for assessing legal compliance and ethical alignment to empower regulators. These novel tools are aimed at systematically evaluating AI systems and providing actionable insights.

B/ Research methodology

Methodology and Data Management

Given the two-fold nature of this project –spurred by significant real-world problems of a practical nature, but deeply grounded in theoretical scientific principles–, we will put forward the following research methodology to ensure that our goal is pursued in a sound and effective way.

From a research perspective, the project will be carried out by international experts guaranteeing the utilisation of state-of-the-art techniques, and ensuring that the execution of the project operates at the forefront of current knowledge and scientific methods. With proven experience in each crucial component of the project (see Section C on the research group's experience and suitability), these experts bring forth not only state-of-the-art methodologies but also the capacity to push boundaries beyond the existing knowledge when necessary.

From a practical perspective, we plan to test our approach in two prominent use cases, i.e., ADS and social assistive robotics. Testing our results in two structurally different scenarios will help ensure that our approach is sufficiently general to be robustly employed in practice.

[5] There have been surveys [Har16, Has18b], EU Projects, such as COMPAS# and SIENNA#, individual initiatives (e.g., https://z-inspection.org/), and some international standards and reports on compliance that are relevant: ISO/IEC 23894 on Artificial Intelligence and Risk Management; ISO/IEC 42001 on Artificial Intelligence — Management System; ISO/IEC 38507 on Governance implications of the use of artificial intelligence by organizations; IEEE P2863 on Recommended Practice for Organizational Governance of Artificial Intelligence; IEEE 7000-2021 on Model Process for Addressing Ethical Concerns During System Design; IEEE 7010-2020 on Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being Draft NISTIR 8332 on Trust and Artificial Intelligence; NIST Special Publication 1270 on A proposal for Identifying and Managing Bias in Artificial Intelligence.

Within each scenario, we will interact with experts who will validate our approach and provide crucial feedback to improve our algorithms and implementations. In the context of such an interaction, we will develop demonstrators with the purpose of (a) incorporating considerations from stakeholders to understand better, criticise, and correct our methodological approaches, but also to (b) better disseminate our scientific results.

Along these lines, we will follow a cyclical approach where we run two rounds of use cases, one at the end of each year to help evaluate and provide feedback for the ongoing work. Despite the short duration of this project, frequent testing in real-life domains can help ensure the impact of our theoretical work and that development is in the right direction.

Next, we describe and motivate our two case studies. Before that, we should stress that our two use cases will focus on two *concrete scenarios* in each application domain. This is important to limit the number of applicable laws and the number of moral values to consider for compliance. Furthermore, concrete scenarios will guarantee focused research and the presentation of results through demonstrators.

Automated driving systems

We shall focus on supporting automated lane-changing decisions in heterogeneous traffic. This is one of the most challenging problems in relation to ADS, will be relevant for many years due to the likely coexistence of ADS and standard vehicles for a long while and it is also relevant in current technology in relation to ADAS, LCDAS and LKAS systems. Besides, per se, it reflects the interplay between legal (with, among others, the EU AI Act and various related ISO and SAE standards) and ethicall (with, among others, the safety vs performance tradeoff and the fact of potentially putting at risk the lives of passengers and pedestrians) that motivate LEGALIS.

Drawing on earlier work by the LEGALIS consortium on decision support in ADS [Cab23], we shall produce an innovative scheme based on adversarial risk analysis to facilitate lane changes to an ADS in presence of heterogeneous traffic. The model will be converted into a simulator that may control for traffic and weather conditions and presence of pedestrians and passengers as well as modify ADS preference and risk aversion parameters. The simulator will be used, on the one hand, as a black box AI to assess its legal and ethical compliance and suggest possible improvements; on the other, we shall use it as a white box AI for the same purpose (since we know the underlying models) to compare results and suggest improvements to the LEGALIS methodology and frame.

Ethical issues: There are a large number of candidate ethical issues that we can consider in this scenario such as how safety is prioritised over other values such as achievement. This is so because individual preferences or damage to material possessions in this context of autonomous driving must be considered less relevant than minimising harm to human life.

Another important ethical issue refers to comparing the safety of passengers vs the safety of nearby pedestrians, see initial ideas in [Cab+22].

Legal issues: There are also multiple legal issues to consider in this scenario. First, there is the need to check conformance with the EU AI Act. Second, there is the need to abide by several ISOs, specially, ISO 17387 on LCDAS (lane change decision aiding systems). Third, there is the need to respect national traffic laws. We will also consider the system accessibility, according to Directive (Eu) 2019/882 Of The European Parliament And Of The Council of 17 April 2019 on the accessibility requirements for products and services (European Assistance Act).

Social assistive robotics

We will draw upon the insights from the ROB-IN project, where an autonomous social assistive robot was deployed in elderly homes, under supervision, to explore personalization and explainability in assistive settings. These essential capabilities were evaluated through fundamental assistive tasks, such as delivering objects to users upon request or issuing personalised reminders. These reminders could be activated either by suggestions from caregivers or by contextual cues—such as reminding a user to drink water if low hydration is detected, encouraging physical activity if the user appears apathetic, or prompting a call to relatives after periods of social inactivity.

Additionally, the robot continuously collects data throughout the day observing the user, which can be potentially communicated to caregivers to enhance the overall care experience or utilised to refine future interactions. The robot employs natural verbal communication facilitated by a sophisticated, yet constrained large language model based on open-source foundational models, enabling effective and intuitive interactions.

Ethical issues: There are a large number of candidate ethical issues that we can consider in this scenario such as disclosure of information, degree of influence of the caregiver directives in the decision-making, robot limiting the elder's autonomy, etc. Additionally, priorities among these ethical issues also need to be considered [Per21, Nor23].

Legal issues: There are also multiple legal issues to consider in this scenario such as data protection, privacy, safety, etc. This case considers a Personal Care Robot (PCR). ISO 13482:2014 'Robots and Robotics Devices – Safety Requirements for Personal Care Robots' applies. We will mainly focus on the levels of autonomy as defined by the Standard: "the robot's capability to execute specific tasks based on current state and sensing without human intervention". There are many normative requirements coming from different sources: EU and national laws, case-based law (judicial decisions), national and EU health and disability policies, and specific constraints set by Hospital ethics committees. For instance, to develop its functionalities, a Care Robot Impact Assessment (CRIA) should be performed . LEGALIS will preliminary single out legal provisions related to the following legal principles already set by the doctrine [Fos15, Fos21]:

- Principle of safety: Prevention of physical/psychical harm.

- Principle of User Protection: Health, Consumer Protection, Environmental Regulation.
- Principle of liability: General liability, Prospective liability, capacity to perform legal transactions and contracts.
- Principle of user rights safeguard: Privacy, data protection, and Intellectual Property Rights.
- Principle of user right to independent living and autonomy: Final say of elderly on the extent to which they receive robot care, enabling human capabilities, acceptance, and persuasion.
- Principle of non-isolation and social connectedness: Non-replacement of human touch and emotions, non-replacement of human caregivers, dignity.
- Principle of autonomous ethical agents' minimization: Limitation to open scenarios with non-mission tasks and human post-monitoring.

Notice that the two use cases that we have selected consider scenarios where AIs are **situated** in the physical world and interact with humans. Hence, both use cases above pose multiple ethical and legal issues that must be compulsorily addressed prior to AI deployment.

In both cases the legal provisions of the GDPR EU Regulation (May 25 2018) and the recent Artificial Intelligence Act (March, 13 2024) apply, in addition to Article 225 of the Treaty on the Functioning of the European Union,the Product Liability Directive 85/374/EEC, Rules 46 and 52 of its Rules of Procedure, the Reports of the Committee on Legal Affairs and the opinions of the Committee on Transport and Tourism, the Committee on Civil Liberties, Justice and Home Affairs, the Committee on Employment and Social Affairs, the Committee on the Environment, Public Health and Food Safety, the Committee on Industry, Research and Energy and the Committee on the Internal Market and Consumer Protection. Especially, risks assessments will be under our consideration. Recital 5 of AIA reads: " AI may generate risks and cause harm to public interests and fundamental rights that are protected by Union law. Such harm might be material or immaterial, including physical, psychological, societal or economic harm".

Expert-driven use case analysis and empirical validation

There is an important aspect that methodologically differentiates LEGALIS from a purely data-driven AI project. Although we propose to develop AI tools for computing legal and ethical compliance, it is the responsibility of human experts to *validate* the results produced by our tools. With that aim, we propose to conduct human-driven *internal* validations and *external* validations.

On the one hand, the research team, as explained below, counts on experts in Applied Ethics (Prof. Begoña Román, UB), Law (Prof. Pompeu Casanovas, IIIA-CSIC), and Responsible AI[6] (Dr. Maria Vanina Martínez, IIIA-CSIC). These internal experts will be tasked with: (i) defining the legal and ethical issues to investigate in each concrete application scenario; and (ii) validating the results from computing legal and ethical compliance within each scenario.

On the other hand, we have arranged an advisory board (see section E) of international experts on "Ethics for Artificial Intelligence", "Law and Technology", "Law, technology and development", "Responsible AI", "Digital economy and Policy-making". These experts will perform an external validation of the results that the LEGALIS AI tools produce.

Data management

Concerning data management, a) in the ADS case study, data will be generated based on a simulator and therefore will provide no fundamental issues as all parametric settings will be under the LEGALIS consortium control; b) In the assistive robotics use case data can come from either simulations or interactions with human volunteers. Experiments involving humans will be conducted in accordance with the European Commission ethics requirements (Human Participants, Personal Data Processing, Incidental Findings Policy, Use of Artificial Intelligence, Misuse of Research Results), European and national legislation and the principles of Responsible Research and Innovation. Depending on the potentially sensitive data to be protected, data management activities will specify the measures to protect and exploit the data, including information notices, consent forms, data sharing, processing agreements and any ancillary data protection documents required to support and regulate project activities; the participation of human volunteers in experiments to collect data will be always supported by the approval of the CSIC Ethical Committee.

Datasets will be available upon request, and after being adequately anonymized, through a FAIR reference site like ZENODO. Software developed will be available through a LEGALIS github site. Publications will be open access.

[6] Dr. Martínez is a member of the United Nations in the Advisory Body on AI. The Body's tasks include building global scientific consensus on risks and challenges, how to harness AI for the Sustainable Development Goals, and how to strengthen international cooperation on AI governance.

C/ Research group's experience and suitability

Research Team Capabilities

The research team is very interdisciplinary, involving researchers on Artificial Intelligence, Law, Applied Ethics, Decision Theory, Statistics and Machine Learning, Robotics, and Human-computer interaction. The research team is composed of three partners: the Spanish National Research Council (CSIC), the University of Verona (UV), and the University of Barcelona (UB). CSIC contributes to the proposal with three research centres: the Artificial Intelligence Research Institute (IIIA-CSIC), the Mathematical Science Institute (ICMAT-CSIC), and the Industrial Robotics Institute (IRI-CSIC). In what follows we describe the academic track record of each partner, previous achievements and innovation capacity as well as proven expertise aligned with the research topic.

PARTNER: The Spanish National Research Council (CSIC)

CSIC is a world-class research institution that provides an excellent research and working environment as well as outstanding networking opportunities. CSIC is the largest public research organisation in Spain[7], ranking fourth in the European Union and sixth in the world[8]. CSIC is under the umbrella of the Ministry of Science and Innovation, though as an independent legal body. The CSIC plays a key role in scientific and technological policy in Spain and worldwide. Its aim[9] is "the promotion, coordination, development and dissemination of multidisciplinary scientific and technological research to contribute to the advancement of knowledge and economic, social and cultural development, as well as the training of personnel and advice to public and private entities in these fields".

The CSIC carries out research, innovation and training in all fields of knowledge – from the most basic or fundamental aspects of Science to the most complex technological developments – distributed in three global areas: Life, Society and Matter. These areas include human and social sciences, food science and technology, biology, biomedicine, physics, chemistry and materials,

[7] SCIMAGO INSTITUTIONS RANKINGS 2022 (Research; Government; EU-28) https://www.scimagoir.com/rankings.php?sector=Government&ranking=Research&country=EU-28
[8] SCIMAGO INSTITUTIONS RANKINGS 2022 (Research; Government; All regions and countries) https://www.scimagoir.com/rankings.php?ranking=Research&sector=Government
[9] Article 4 of CSIC's Statute - https://www.boe.es/buscar/act.php?id=BOE-A-2008-591

natural resources or agricultural sciences, among others. For this purpose, it employs almost 4,000 researchers, distributed in its 121 research institutes[10] across the country.

The CSIC leads the scientific production of Spain, with an annual average of 13,000 publications in internationally renowned scientific journals, with a very high percentage of publications in frontline journals: more than 70 % of the total published articles correspond to high impact articles (Q1). Regarding Knowledge Transfer, the CSIC is the top institution in Spain in patent generation, with 85 patent applications in 2021[11].

CSIC has a broad experience in conducting R&D projects funded by national and international public agencies and industry. CSIC excels at research and also at attracting funding and at hosting MSCA and ERC international researchers. It is a major player in the development of the European Research Area, and therefore a significant contributor to the European integration process. In Horizon 2020, the EU R&I Framework Programme for 2014-2020, the CSIC was ranked as the first organisation in Spain and third in Europe by the number of actions. During H2020, the CSIC was granted a total of 891 projects, coordinating 82 of them, and with a total EU financial contribution of EUR 382 million. In Horizon Europe, the CSIC achieved a total of 98 projects as of February 2022, 5 of which were coordinated. The CSIC is also a leading player in the ERC programme, with more than 120 projects signed as a Host Institution.

The Artificial Intelligence Research Institute (IIIA-CSIC)

In this proposal, the CSIC's Artificial Intelligence Research Institute (IIIA-CSIC) will act as coordinator. IIIA-CSIC is a publicly-funded research, being one of Spain's top research centres on Artificial Intelligence. The IIIA-CSIC undertakes basic and applied research on machine learning, multi-agent systems, and reasoning and logic. Founded in 1994, it currently has 87 full-time staff, including 25 permanent senior researchers, six post-doc researchers, 30 Ph.D. students, and 7 Master students. The centre has a dedicated unit for knowledge and tech transfer activities (UDT-IA). The IIIA-CSIC is "Grup de Recerca Consolidat (SGR)" and has the TECNIO recognition for its innovative technologies and technology transfer activities, both recognized by the Generalitat de Catalunya (Catalan Government). The IIIA is currently the coordinator of the CSIC's Artificial Intelligence Hub, which bring together more than 400 researchers working in fundamental AI and applied AI to more than 19 scientific disciplines, including material science, nanotechnologies, physic, astrophysic, ocean, climate, neurosciences, biology and bioinformatics, education, health, etc.

[10] https://www.csic.es/es/investigacion/institutos-centros-y-unidades

[11] https://www.oepm.es/es/sobre_oepm/noticias/2022/2022_04_05_Record_solicitudes_patente_europea.html

IIIA-CSIC´s researchers play leadership roles in the European research communities of machine learning, agent technologies, and reasoning. The IIIA-CSIC is the European centre with more Fellows of the European Association of AI, EurAI, currently seven. Prof. Juan A. Rodríguez-Aguilar, the PI of this proposal, is one of them. Besides that, IIIA researchers have received many important international awards in Artificial Intelligence. The most recent ones are the ACM/SIGAI Autonomous Agents Research Award 2019 to Prof. Carles Sierra, the SEIO-Fundación BBVA 2021 "Mejor contribución metodológica en Investigación Operativa" to Christian Blum, and Premios de Investigación Sociedad Científica Informática de España (SCIE) – Fundación BBVA 2023 "Jóvenes investigadores", to Marc Serramià Amorós, former PhD student of Prof. Juan A. Rodríguez Aguilar, the PI of this proposal. Currently, the IIIA-CSIC's director is the President of the European Association of AI, EurAI.

IIIA-CSIC has produced close to 495 publications in the last five years, including journal articles (230), conference and workshop articles (176), and books (89). IIIA's publications are mostly on fundamental results in AI. Nonetheless, we also publish in some of the most well-known journals in applied AI. Around 50% of our publications go into Q1 journals. During the last 5 years the IIIA-CSIC has obtained 48 research projects (17 with EU funding, 14 with national funding) in competitive calls, of which 34 are currently ongoing. There are 10 ongoing European projects. IIIA-CSIC coordinates "Value Aware Artificial Intelligence" project, VALAWAI (HORIZON-EIC-2021-PATHFINDER CHALLENGES-01-01), and an MSCA COFUND "Artificial Intelligence in Development Goals", ALLIES (HORIZON-MSCA-2022-COFUND-01-01). IIIA-CSIC participates in two training networks: NL4XAI (H2020-MSCA-ITN-2019-860621) and EUROVA (H2020-MSCA-ITN-2019-860960) on explainable AI and machine learning for health, respectively. The IIIA-CSIC is also participated in several European Commission's strategic projects on AI: "A European AI On-Demand Platform and Ecosystem" (AI4EU H2020-ICT-26-2018-825619), "Towards a vibrant European network of AI excellence centres" (Human-AI-Net H2020-ICT-2019-3-952026), and the Foundations of Trustworthy AI (TAILOR H2020-ICT-2019-3-952215).

Expertise. The IIIA-CSIC research team has a strong position on AI & Ethics and AI & Law. The IIIA-CSIC team contributes to the project with experts on *AI & Ethics* (Dr. Pablo Noriega, Dr. Nardine Osman, Prof. Juan A. Rodríguez-Aguilar), *AI & Law* (Prof. Pompeu Casanovas, Dr. Pablo Noriega), *Norms (regulation) in Artificial Intelligence* (Dr. Pablo Noriega, Dr. Nardine Osman, Prof. Juan A. Rodríguez-Aguilar), *Natural Language Processing* (Dr. Nardine Osman), *Machine Learning* (Dr. Filippo Bistaffa, Dr. Nardine Osman, Prof. Juan A. Rodríguez-Aguilar), Interpretability and Explainability (Dr. Nardine Osman, Dr. Maria Vanina Martínez), and Formal languages and Formal verification (Dr. Maria Vanina Martínez, Prof. Juan A. Rodríguez-Aguilar).

Dr. Nardine Osman leads the VAE[12] project (TED2021-131295B-C31) and is the scientific coordinator of the Pathfinder VALAWAI[13] project. These projects focus on value engineering and on building AIs that can manage ethical values. Thus, they are both tightly related to this proposal. Prof. Juan A. Rodríguez-Aguilar currently leads (as co-PI) the ACISUD[14] project (PID2022-136787NB-I00), which focuses on developing AI for the social good. There, he leads work on building ethical autonomous cars. Furthermore, Dr. Filippo Bistaffa and Prof. Juan A. Rodríguez-Aguilar lead the YOMA OR[15] project (OPE02570), where they investigate how to exploit Generative AI to recommend learning pathways to young learners in Africa. Therefore, ACISUD and YOMA OR are tightly related to this proposal. Dr. Martínez is currently the PI of the DESINFSOC[16] project (PIE 20235AT010) and is also part of the research team of the LYNEXSYS[17] project (PID2022-139835NB-C21) both projects focus on the combination of state-of-the-art ML and NLP techniques with symbolic approaches towards Explainable Intelligent Systems. Dr. Pablo Noriega leads the DESAFIA2030 project (BILTC22005) whose purpose is the design of an educational strategy and accompanying follow-up actions on the ethical challenges of AI. In the past, Dr. Martínez was the scientific coordinator for Argentina in the project DODAM[18], on the development of declarative and ontology-enhanced data analytics and machine learning methods, funded by the scheme STIC AMSUD and participated on the research team of the EU H2020 U H2020 Marie Sklodowska-Curie project "MIREL: MIning and REasoning with Legal texts[19]" (grant agreement No. 690974) specifically working on the use of computational argumentation frameworks for the representation of and reasoning about legal knowledge.

Prof. Casanovas leads the ethical and legal WP9 of the EU H2020 OPTIMAI[20] project (Optimising Manufacturing Processes through Artificial Intelligence and Virtualization, ID: 958264). He is building with Dr. Mustafa Hashmi the OPTIMAI regulatory model and legal ecosystem. With this background, Prof. Casanovas and Dr. Noriega will be mainly involved in legal knowledge representation (RO1 and WP3) and validation tasks involving the demonstrators (TO2 and WP6); Dr. Osman will be involved in ethical knowledge representation (RO1 and WP3) and the computation of ethical compliance (RO3 and WP4); Dr. Bistaffa and Prof. Rodríguez-Aguilar will be involved in exploiting generative AI to retrieve applicable laws (RO1 and WP3) and in the computation of legal compliance (RO3 and WP4). Dr. Martínez will contribute to explaining an AI to stakeholders (RO4 and WP5); and Dr. Bistaffa and Prof. Rodríguez-Aguilar will lead the development of the LEGALIS demonstrator.

[12] https://www.iiia.csic.es/en-us/research/project/?project_id=238
[13] https://valawai.eu/
[14] https://www.iiia.csic.es/en-us/research/project/?project_id=246
[15] https://www.iiia.csic.es/en-us/research/project/?project_id=242#description
[16] https://www.iiia.csic.es/en-us/research/project/?project_id=247
[17] https://www.iiia.csic.es/en-us/research/project/?project_id=244
[18] https://www.sticmathamsud.org/sitio/wp-content/uploads/2022/01/STIC-AmSud-Projects-2021-call.pdf
[19] https://cordis.europa.eu/project/id/690974/es
[20] https://optimai.eu/

PI excellence. Prof. Juan A. Rodríguez-Aguilar is a Research Professor at the Department of Multiagent Systems at IIIA-CSIC. He has led 14 projects as PI, including international, European, national, and industrial projects. He has also participated in 26 research projects as team member. He has published over 250 articles, receiving nearly 7800 citations (H-index=42) in his 25 years of experience in AI. He is an expert in optimisation and formal verification of multi-agent systems, and trustworthy AI, two research fields that are core in this proposal. Moreover, he has been teaching formal verification techniques at the Menendez Pelayo International University since 2016, and he has collaborated with researchers at CERN to improve their verification techniques. He has mentored 12 PhD students, 16 M.Sc. students, and one post-doc (formerly supervised as an MSCA fellow). His worldwide collaborations (University of Verona, University of Oxford, Centre National de la Recherche Scientifique, Vrije Universiteit of Brussels) are evidenced by joint publications. He is also recipient of several awards: the European Association for Artificial Intelligence Fellowship (2013), the most cited article of the Engineering Applications of Artificial Intelligence Journal, the Best Agent Application of the EU-funded Agentcities International Competition, and Distinguished SPC member of the IJCAI conference (2018, 2022). He is a member of the editorial board of top AI journals (e.g., AI Journal, Journal of Artificial Intelligence Research, Autonomous Agents and Multiagent Systems). In 2024 he was elected board member of the International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS). He regularly acts as an expert for international research agencies (e.g., UK, US, Portugal), including the EU Commission's. Prof. Rodríguez-Aguilar has ample experience on the industrial application of AI techniques and on technology transfer. In 1999 he was one of the founders of iSOCO S.A., the first spin-off of the CSIC, acquired in 2016 by Clever Global, a publicly-traded company. The CSIC has registered the intellectual property of 6 of the software developments resulting from his research.

The Institute of Mathematical Sciences (ICMAT-CSIC)

Another CSIC participating institute will be ICMAT (the Institute of Mathematical Sciences) a joint research centre from CSIC and three main universities in Madrid: UAM, UC3M and UCM. Created in 2007, it is located in a modern building with outstanding research facilities. ICMAT hosts a strong core of excellent and highly-influential researchers, who have contributed groundbreaking results that have solved a wide range of long-standing open problems and conjectures. Notable examples of its impact are its impressive array of publications in the most reputed mathematical and statistical journals, the 12 ERC awards given to ICMAT researchers (50% of Spain's ERC awards on the PE1 panel), one AXA Chair and 4 members of the Spanish Royal Academy of Sciences. ICMAT is **one of the only two research centres in Spain** to have been awarded **four consecutive Severo Ochoa** (SO) awards (2011, 2015, 2019 and 2024). Its Computing Services Center (CSC) provides support for scientific computing at ICMAT. The main ICMAT cluster, "LOVELACE," offers high computing capacity, consisting of more than 900 physical cores, specialised hardware (XeonPhi, 5xNVIDIA V100, 3xNVIDIA A100), as well as a high-speed and fault-tolerant Lustre file

system that can support both serial and massively parallel workloads, well-suited for the computations required in LEGALIS.

Expertise

ICMAT will contribute to LEGALIS with expertise in Decision Theory, Bayesian Statistics, Risk Analysis, Causal Inference and Probabilistic Machine Learning and their applications in various AI domains including Automated Driving Systems through its DataLab https://datalab.icmat.es/.

It is led by David Rios Insua who is Research Professor at ICMAT-CSIC and member of the Spanish Royal Academy of Sciences. He was AXA-ICMAT Chair in Adversarial Risk Analysis until last December. He is author of 20+ books, 160+ refereed papers and book chapters, in his areas of interest which include decision analysis, risk analysis, Bayesian statistics, machine learning and negotiation analysis, and their applications, mainly, to safety, security, and cybersecurity. He has been PI in 60+ sponsored projects totaling more than 7.5 M€. He has received national research awards from SEIO (2), Everis Foundation, FEI and Wirsbo and international research awards from IIASA (Peccei), INFORMS (Edelman), SRA, DAS and ISBA (DeGroot). He has codirected two NSF-SAMSI programs and coined the term and methodology of Adversarial Risk Analysis. He directed scientifically the H2020 CYBECO project on cybersecurity and cyber insurance and has participated in the H2020 projects TRUSTONOMY (Trust in ADS) and STARLIGHT (Secure ML) of direct relevance to LEGALIS. He has held visiting positions in: Austria IIASA; USA: Duke, Purdue, SAMSI; Italy: CNR-IMATI; France: PSL-Université Paris-Dauphine; Finland, Aaalto; PRC, USST. He has had 19 contracts with administrations and companies totaling 7,35 M€. A major achievement is leading the AESA-RAC project developing a methodology for aviation safety risk management leading to annual savings of 800M€. He has been scientific director of Aisoy Robotics which has sold over 3000 social robots globally. His research has led to AI systems available at AESA, Aisoy, Xeerpa, A3sec and Intrasoft. He is a member of the Higher Statistical Council in Spain, where he led the group on data governance. He has supervised 27 PhD theses and is on the board of the CSIC AIHUB-Connection.

He will be assisted by a group of motivated PhD students which include Carlos G. Meixide (causal inference), José M. Camacho (cybersecurity in AI), Pablo G. Arce (secure machine learning) and Pablo Varas (Transformer architectures for scientific discovery), as well as postdocs, including Roi Naveiro and Victor Gallego and two postdocs to join through the COFUND-ALLIES program. The Lab has also an extensive network of collaborators at Duke, Aalto, CNR-IMATI, CNRS-LAMSADE, IBM Research and UT Austin, to name but a few, in lines related to LEGALIS.

The ICMAT team will link with the other LEGALIS partners especially in connection to RO3 (adversarial machine learning aspects), RO4 (causal inference and sensitivity analysis), and TO1

(in relation to the ADS case study), contributing to tasks T1.1-2, T3.4, T4.1-3, T5.1-2, T6.1-3, as specified in the work plan below.

The Institut de Robòtica i Informàtica Industrial (IRI-CSIC)

The Institut de Robòtica i Informàtica Industrial (IRI-CSIC) is a key player in the Spanish robotics and automatic control scenes, and a valued participant in a large number of international collaborations, including the prestigious European networks EuRobotics, CLAIRE or ELLIS. IRI-CSIC received the María de Maeztu scientific excellence seal for the period July 2017 to June 2021. In this period, a strategic research program in human-centred robotics was developed. This is the main accreditation given by the Spanish Government to research units that stand out for the impact and international relevance of their results.

IRI-CSIC is very active and successful in European, national and regional competitive research programmes. For instance, it leads or participates in various enabling research projects for assistive robotics: ROB-IN, focusing on personalization and explainability of robots in domestic care settings; CLOE-GRAPH, on high-level task representation and explainability; SeCuRoPS, addressing privacy and security in social robots; and TRAIL, which centres on transparency in human-robot interactions. Additionally, IRI-CSIC spearheads the Open Laboratory for Assistive Robotics.

The IRI-CSIC research contributes to LEGALIS with its extensive experience in social robotics (Dr. Guillem Alenyà), explainability in social robotics including counterfactual explanations (Dr. Guillem Alenyà, Msc Tamlin Love), and ethics in AI (Dr. Cristian Barrué). Dr. Cristian Barrué is member of The Observatory on Society and Artificial Intelligence of the European platform of AI[21]. The team is working closely with the Catalan Observatory of Ethics in AI to develop specific guidelines and evaluation tools for roboethics.

PARTNER: University of Barcelona (UB)

The University of Barcelona is highly placed on a national, European or global scale in the most prestigious academic rankings, which reflect the assessment of indicators such as teaching and research quality, knowledge transfer and internationalisation. In the Academic Ranking of World Universities (ARWU) it ranks number 1 at national level, and 201 at global level. In the Best Global Universities Rankings it ranks number 1 at national level, 27 at European Level and 86 at global level.

Within UB, the research group Language and Computation Centre (CLiC) focuses its research on the computational processing of language. CLiC has been recognized as an established group by Generalitat de Catalunya (2021 SGR 00313) since 2015. CLiC consists of 12 senior researchers,

[21] https://www.ai4europe.eu/ethics/osai

from which 3 participate in this project. Additionally, the Aporia research group develops its activity within the framework of Contemporary Philosophy and Ethics. It has also been recognized as an established group by Generalitat de Catalunya (2021 SGR 00294). This group has 20 members, and the project counts with one of the two coordinators.

The CLiC group contributes to the project with experts in the computational characterisation of Ethical Values and Normative Multi-Agent Systems (Prof. Maite López-Sánchez [Rod23, Ser23, Rod22, Rod21, Rod20]), User-eXperience, Human-Computer Interaction and chatbots (Dr. Inmaculada Rodríguez-Santiago [Kav23a, Kav22, Tel20]), as well as Visualisation for Explainability and Natural Language Processing (Dr. Anna Puig [Kav24, Kav23b, Zou22]). Additionally, Dr. Begoña Román [Par21, Par20, Roy15] is an expert on Ethics and coordinates the Aporia group. Prof. López-Sánchez and Dr. Román will actively collaborate on representing ethical knowledge to contribute to RO1, TO2, WP3 and WP6. Dr. Rodríguez-Santiago and Dr. Puig will mostly contribute to RO4, TO1, and WP5.

Here we highlight five of the most relevant research projects these researchers are actively involved in. GRAPES: learninG, pRocessing, And oPtimising shapES[22]. VAE: Value-Awareness Engineering[23]. FairTransNLP-Language: Analysing toxicity and stereotypes in language for unbiased, fair and transparent systems[24]. The Philosophical view as a medical view[25]. ACISUD: Advanced Computational Intelligence Techniques Used for the Benefit of Sustainable Development[26].

PARTNER: University of Verona (UV)

The Department of Computer Science of Verona University, was established in 2001 and promotes numerous teaching and research activities based on a continuous exchange of skills ranging from mathematics to physics through computer science seen as science and engineering. Approximately 200 people collaborate in the Department of Computer Science, including teachers, researchers, doctoral students and technical administrative staff. In 2016, the Ministry of University and Research (MIUR) rated the Department of Computer Science of the University of Verona among the top 5 best places in Italy to learn and apply innovation in Computer Science. In 2018, the National Agency for the Evaluation of the University and

[22] Project funded by the European Commission. Project info: https://cordis.europa.eu/project/id/860843.
[23] Project funded by Ministerio Ciencia e Innovación. Project info: https://www.iiia.csic.es/es/research/project/?project_id=238#description
[24] Project funded by Ministerio de Ciencia e Innovación (MICINN). Project info: https://clic.ub.edu/en/node/597
[25] Project funded by Ministerio de Economía y Competitividad. Project info: https://www.ub.edu/grc_aporia/en/projects/project-the-philosophical-view-as-a-medical-view/
[26] Project funded by Ministerio de Ciencia e Innovación. Project info: https://www.csic.es/en/research/research-projects/advanced-computational-intelligence-techniques-reaching-sustainable-development-goals

Research System (ANVUR) awarded and financed the Department of Computer Science of the University of Verona as a "Department of Excellence", electing it among the best locations for carrying out Industrial Innovation in Italy. The mission of the Department of Computer Science has always been to promote excellence in scientific research in its areas of expertise as well as offer its students cutting-edge teaching in step with the needs of the constantly evolving market. This is demonstrated by the numerous scientific production (over 2000 international publications since 2018), the participation in projects financed by international and national bodies (about 20 projects for about 10 million euros) as well as the collaboration activity with companies through applied research with a strong technological impact (7 spin-offs and newly established companies, 10 patents, and more than 150 innovation projects for about 5 million euros).

The research activities related to this project will be carried out mainly by people involved in the Intelligent System Lab (ISLa). Over the last ten years, the Intelligent System Lab (ISLa) has been developing methodological and applied research in artificial intelligence, machine learning and data analysis for intelligent systems. The methodologies studied range from reinforcement learning and planning with uncertainty, to multi-agent coordination, probabilistic modelling and statistical data analysis. Application domains range from industry 4.0, to security and cyber-physical systems, to name a few examples. The group participates in national and international research projects and has collaborations with Italian and foreign research centres. The group currently has more than ten people including teachers, PhD students and research collaborators.

The ISLa group contributes to the project with experts in Safe Reinforcement Learning and Multi-Agents Systems (Prof. Alessandro Farinelli [Ami+23] [Bis+21] [Mar+24]) and Neuro Symbolic AI (Dr. Daniele Meli [Mel+21] [Ver+23] [Mel+24]). Both Prof. Farinelli and Dr. Meli will actively collaborate on learning an interpretable model of AI from observations to contribute to RO2, TO1 and WP2.

Here we mention three of the most relevant research projects these researchers are and have been actively involved in Development and application of Novel, Integrated Tools for monitoring and managing Catchments (INTCATCH, https://intcatch.eu/); Computer Engineering for Industry 4.0 (https://www.di.univr.it/?ent=progetto&id=4935); Autonomous Robotic Surgery (ARS, https://www.di.univr.it/?ent=progetto&id=4831)

External collaborators

Besides the three institutions composing the research team, we will count on three renowned external collaborators to assist in the preparation of the legal repositories that we will need to help legal experts retrieve applicable laws. These experts are Dr. Víctor Rodríguez Doncel (Polytechnical University of Madrid), Dr. Mustafa Hashmi (University of La Trobe, Australia), and Prof. Brian (Ho-Pun) Lam (Xi'an Jiatong-Liverpool University, China) . The three of them are close collaborators of Prof. Pompeu Casanovas' (IIIA-CSIC). Dr. Víctor Rodríguez Doncel, Dr. Mustafa Hashmi, and Brian (Ho-Pun) Lam will contribute to representing legal knowledge (RO1), and in

particular to arranging the legal repositories that we will employ as external knowledge for the LLM to be employed by legal experts (see task T3.1 in the work plan below). On the one hand, Dr. Víctor Rodríguez Doncel is an expert on semantic representation, legal ontologies and (legal) knowledge graphs. On the other hand, Dr. Mustafa Hashmi and Prof. Brian (Ho-Pun) Lam are both experts on business and legal compliance modelling, and rule extraction from normative legal provisions. Their expertise will be invaluable to guarantee the success of retrieving applicable laws and to formalise executable rules (from norms).

Proposed governance and collaboration approach

IIIA-CSIC, ICMAT-CSIC and IRI-CISC belong to CSIC's Artificial Intelligence Connection AIHUB[27]. This network of CSIC research centres brings together more than 400 researchers from more than 80 research groups from 40 centres that apply AI in research areas as diverse as artificial intelligence, physics, mathematics, robotics, microelectronics, life sciences, astronomy or philosophy. AIHUB is a network coordinated from IIIA-CSIC and IRI-CSIC that seeks to promote the connection of these groups with each other and with society through regular meetings and joint projects. One of the results of the network is an ongoing European project: the MSCA COFUND "Artificial Intelligence in Development Goals", ALLIES (HORIZON-MSCA-2022-COFUND-01-01). IIIA-CSIC, ICMAT-CSIC and IRI-CISC are partners, and collaborators, in the ALLIES project. Therefore, this PRISMA proposal would serve to strengthen the collaboration between the three groups.

IIIA-CSIC and UB have a long history of successful collaborations through joint projects (e.g., the ongoing ACISUD[28] (PID2022-136787NB-I00), CI-SUSTAIN[29] (PID2019-104156GB-I00), H2020 CROWD4SDG[30] (https://cordis.europa.eu/project/id/872944), to name a few), co-supervision of PhD students and MSc students, and research publications (e.g., [Rod20, Rod21, Rod22, Rod23], to name only a few that are related to this proposal).

IIIA-CSIC and UV have also a long history of successful collaborations. Prof. Farinelli and Prof. Rodríguez-Aguilar, both taking part in this proposal, started publishing together back in 2010. Since then, they have co-authored multiple Q1 journals and high-impact conference papers. The most recent article resulting from the collaboration of the two groups [Ver+23] is tightly related to this proposal. Furthermore, there is a long history of exchanges of PhD students and postdocs. For instance, Dr. Filippo Bistaffa, part of the research team, obtained his PhD in Verona, visited

[27] https://aihub.csic.es/en/
[28] https://www.iiia.csic.es/en-us/research/project/?project_id=246
[29] https://iiia.csic.es/en-us/research/project/?project_id=211
[30] https://cordis.europa.eu/project/id/872944

IIIA-CSIC as a PhD student, enjoyed a Marie Curie at IIIA-CSIC, and he is now a tenured scientist at IIIA-CSIC.

Prof. Pompeu Casanovas and Pablo Noriega have been cooperating in large national (White Book on Mediation[31]) and EU projects (AT[32], SINTELNET[33]) and publications for more than twenty years now. They published their first book together in 2007 [CAS07]. For the last three years they have been working on dilemmas in legal governance [Cas21], AI governance [Nor22], and regulation of complex systems [Cas23]. Likewise, the collaboration between Prof. Casanovas' research group (IDT-UAB) and Prof. Rodríguez-Doncel's Ontology Engineering Group (UPM) dates back to the EU Project SEKT (2003-2006)[34]. LYNX[35], led by Profs. Rodríguez-Doncel and Elena Montiel, lasted from 2017 to 2021 on the building of a EU legal knowledge graph for smart compliance services. They also have a long history of books, articles and chapters in common, e.g., the volumes *Linked Democracy* [Pob19] and AICOL [RodD21], and [RodD16, Cas17]. It is worth mentioning here the Law, Science and Technology EU Doctorates, the Erasmus LAST (2012-2016)[36], and the Marie-Curie LAST-RIoE (2019-2023)[37], in which the three of them have been cooperating in a joint endeavour since 2012. Dr. Mustafa Hashmi and Prof. Brian (Ho-Pun) Lam were experts on business language models at CSIRO, where they worked on business language models, RuleML and compliance with Prof. Guido Governatori. Prof. Lam developed the reasoner SPINdle. From 2016 onwards, in occasion of the Australian Cooperative Research Centres (CRC) Projects Data2Decisions[38] and iMove[39], Prof. Governatori and Dr. Hashmi joined the La Trobe Lawtech Research Group. Prof. Lam also participated and fully cooperated in meetings and publications with the team. Since then, Dr. Hashmi and Prof. Lam have been publishing extensively on compliance in the legal field [Hash18b, Cas2024a,b,c]. It is worth mentioning, because it is relevant for this project, the La Trobe Report (2021) on the Proposed regulatory approach to recognise Connected and Automated Vehicles in the Disability Standards for Accessible Public Transport 2002 (iMove).

To summarise, the partners of this project have a strong history of collaboration, with exceptional research results. The network of personal relations is also very good amongst the partners. This should have a significant positive impact in management, where we foresee few conflicts, and in case they do arise, we foresee a good atmosphere for resolving them.

The IIIA-CSIC will serve as coordinator in the project and will take care of governance. IIIA-CSIC will take care of the everyday management of the WPs and their tasks to ensure their proper and

[31] https://ddd.uab.cat/pub/llibres/2010/168589/libro_blanco_mediacion_a2010iSPA.pdf
[32] https://www.agreement-technologies.eu/
[33] https://cordis.europa.eu/project/id/286370
[34] https://www.sekt-project.com/ , https://cordis.europa.eu/project/id/506826
[35] https://lynx-project.eu/data2 , https://cordis.europa.eu/project/id/780602/es
[36] http://www.last-jd.eu/
[37] https://cordis.europa.eu/project/id/814177
[38] https://www.d2dcrc.com.au/
[39] https://imoveaustralia.com/

timely execution and the fulfilment of the project objectives. Furthermore, as discussed above, the partners of this project have a strong history of collaboration, with exceptional research results. The network of personal relations is also very good amongst the partners. This should have a significant positive impact in management, where we foresee few conflicts, and in case they do arise, we foresee a good atmosphere for resolving them.

We plan to have a physical kick-off meeting at the beginning of the project (M1) to plan our work, and an online closing meeting at the end (M24) to plan future work and funding. During the project, we will have three main consortium meetings every six months, at months M6, M12, and M18. Meetings will take place online to reduce travelling. In addition to the consortium meetings, which will be 1–2 full day meetings, we will also have short frequent meetings every two months to ensure the proper and timely execution of our work. Note that the consortium and advisory board meetings will also be used to verify our milestones as specified in detail in the work plan (see section F below). In addition to these general project meetings, there will also be focused meetings, online or physical, as deemed necessary, for the researchers collaborating on similar topics. These focused meetings will be planned as needed during the project meetings.

D/ Expected results and applicability

Project Impact

In a globalised world, the adoption by businesses and society of intelligent computing systems is vital to ensure a competitive and sustainable society. Accordingly, the European Digital Strategy and the Spain Digital Agenda 2025 aim to accelerate a humanistic digital transition in our country. However, some factors still prevent the massive, secure, and reliable adoption of AI. This reluctance is primarily a product of the fact that interactions between people and AI systems are limited by a lack of understanding of AI systems and the lack of tools for assessing their ethical and legal compliance before they are deployed. Resistance is driven by automated decisions that are difficult to understand, legal risks, or automated solutions that are not aligned with ethical values.

LEGALIS will contribute to the adoption of AI by producing a body of knowledge and technologies that will pave the way for the practical implementation of regulatory sandboxes [Mad+22, Oec23], and the assessment of ethical compliance, removing the barriers to the adoption of AI by businesses and society in general and, thus, fostering the digital transition. Furthermore, the project aims to contribute to the area of humanistic or value-based AI in the EU.

More specifically, LEGALIS has potential to have an impact on enabling the practical deployment of automated driving systems and social assistive robots.

Our first use case focuses on the ethical and legal aspects of ADS. ADS have emerged as a potential solution to modern day transport problems. Making cities inclusive, safe, resilient, and sustainable is one of the goals set out by the United Nations (UN) in their 2030 Sustainable Development Goals (SDGs). Sustainable transport systems and infrastructure can reduce the negative impacts of urban development on the environment, economy, and society [Gop+15]. Widespread AV adoption can reduce environmental degradation through reduced emissions and energy consumption while providing beneficial economic and social outcomes through improved efficiency, traffic flow, road safety, and accessibility to transport, among other benefits [Pet+15, Dom+16]. Much of these benefits stem from AVs' connected nature, which enables them to communicate with other vehicles and critical infrastructure to optimise traffic and maximise all associated benefits for sustainable and smart cities [Cab23]. Furthermore, as driving automation takes over the difficult and tiring parts of driving, driver comfort and safety will increase. These benefits will likely extend beyond private vehicles, supporting new shared mobility opportunities – such as robotaxis, with the potential to reduce congestion and CO2 emissions, as the research team has already investigated [Bis+21]. However, widespread adoption of ADS will not take place as long as users fully trust them, with such trust relying on users understanding ADS decision making, their compliance with legal systems and sharing their ethical alignment. Therefore, we argue that LEGALIS has potential to benefit the public administration (regulators and policymakers) and industry (car makers) by providing them with tools to evaluate ethical and legal compliance. Ultimately, it will also benefit citizens acquiring AVs or using their services.

Regarding our second use case, notice that we will address the new care economy. LEGALIS is very timely because the assistive robotics industry is now developing and still it is not clear how to enforce current legislation, both in terms of legality and ethics. There are several initiatives to deploy robots in various spaces like hospitals [Mor22], retirement homes, and even private homes (ROB-IN project) that usually focus only on the technical aspects. By developing the second use case, our objective is to guarantee the legal and ethical compliance of assistive robots, whose deployment will aid to improve healthcare assistants' productivity and public health services. Likewise for our first use case, there is potential for LEGALIS to benefit the public administration and industry, as well as patients.

Another significant benefit of the project is in terms of talent generation and employment. The project will hire four research engineers. Their learning process during this project will open the door to future employability in the R&D ecosystem or technological companies.

We propose to assemble an advisory board whose members are very related to the two use cases we will research. Furthermore, ICMAT-CSIC has important experience in researching ADS (first use case), and IRI-CSIC has also a successful track record in developing social assistive robots (second use case). The consortium will capitalise on the networks of the advisory board and two of our team members, ICMAT-CSIC and IRI-CSIC, to organise meetings with stakeholders to introduce the project results (particularly the software demonstrator) and foster the co-creation of novel products and services.

Furthermore, we will approach the public administration and technological companies, preferably delivering legal services, in the related areas to validate the applicability of the use case demonstrators. These activities with the public administration and companies, and other stakeholders aim to explore potential technology transfer actions. The consortium will study different intellectual property protection methods to support the successful development of tech transfer activities if there is room for it. We plan to explore at least two software registration and patenting studies of the most promising and innovative results regarding ethical and legal compliance respectively.

E/ Work plan and schedule

Detailed steps of implementation: organised in work packages

To work towards the above-detailed global objective, we propose a research agenda structured into five work packages (plus 1 management work package). Figure 1 provides a graphical overview of the LEGALIS project, its WPs, and their interactions.

WP1 focuses on management and maximising impact (M1–M24). WP2 focuses on exploiting AI techniques to learn a model of an AI to inspect by external observation. WP3 is concerned with developing AI tools for helping experts in their knowledge engineering tasks (retrieve applicable norms and define moral values and value systems). WP4 develops AI tools for computing and assessing legal and ethical compliance and WP5 focuses on developing AI tools for building explanations. Finally, WP6 wraps up the developments in the former work packages into a demonstrator that will serve to showcase the results of the project to relevant stakeholders. The Gantt diagram in Figure 3 schedules each work package and its tasks.

In what follows we present the description of the work packages, followed by a list of deliverables (Table 1), and a list of milestones (Table 2). Furthermore, Table 3 identifies risks together with risk-mitigation measures we plan to undertake if needed.

WP1	**Management and impact creation**	**M1**	**M24**
Leader	**IIIA-CSIC**		
Aim of the WP			

● To ensure planning, implementation, coordination and achievement of the goals ● To ensure timely completion of milestones and successful completion of the work package tasks ● To assist decision making, internal and external communications and dissemination, encourage greater accountability and control, minimise risk, identify and exploit project related opportunities ● To help partners achieve their project objectives. The project management comprises several issues including technical, administrative, financial, communication and knowledge management issues inside the project, and external relationships between the project and BBVA Foundation.	
Tasks	
T1.1	**General management (M1-M24: responsible IIIA-CSIC; involved ICMAT-CSIC, IRI-CSIC, UB, UV).** This task is concerned with the everyday management of the WPs and their tasks to ensure their proper and timely execution and the fulfilment of the project objectives. Furthermore, as discussed above, the partners of this project have a strong history of collaboration, with exceptional research results. The network of personal relations is also very good amongst the partners. This should have a significant positive impact in management, where we foresee few conflicts, and in case they do arise, we foresee a good atmosphere for resolving them. We plan to have a physical kick-off meeting at the beginning of the project (M1) to plan our work, and an online closing meeting at the end (M24) to plan future work and funding. During the project, we will have three main consortium meetings every six months, at months M6, M12, and M18. Meetings will take place online to reduce travelling. In addition to the consortium meetings, which will be 1–2 full day meetings, we will also have short frequent meetings every two months to ensure the proper and timely execution of our work. Note that the consortium and advisory board meetings will also be used to verify our milestones. A summary of these meetings is presented in Figure 2 below. In addition to these general project meetings, there will also be focused meetings, online or physical, as deemed necessary, for the researchers collaborating on similar topics. These focused meetings will be planned as needed during the project meetings.
T1.2	**Dissemination and communication (M1-M24: responsible IIIA-CSIC ; involved ICMAT-CSIC, IRI-CSIC, UB, UV).** This task is very important to achieve the social and economical impact expected in the project. It is oriented to the spread of knowledge the project will deliver to the research community and to create the relationships with different stakeholders to showcase and look for opportunities for technology transfer. Its main activities, which we list below, are defined to achieve impact.

	• Coordinate general communication activities internally and with the consortium communication departments. Develop a communication plan including the development of communication materials (news and press releases) related to the project activities and results. Communicate the messages through different channels and networks. • Scientific publications: Each sub-project leader will organise a plan for publications. The dissemination and exploitation plan below already identifies relevant, top-indexed journals and high-impact conferences. • Organisation of meetings with stakeholders: Identify contacts, coordinate and execute meetings related with the execution, evaluation and demonstration of the use cases. These meetings will be oriented to pursue new collaborations and agreements. • IP Protection: Identify the most promising results that can be subject of protection, achieve the patenting analysis and implement the software registration.

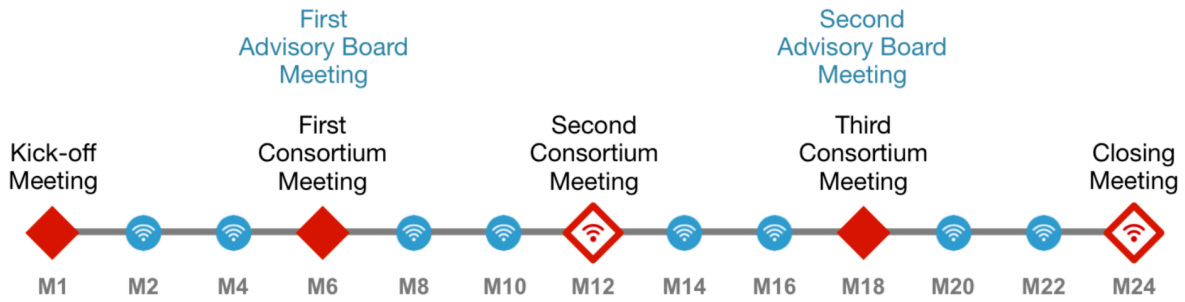

Figure 2. Project meetings.

WP2	Model learning from observation	M1	M24
Leader	UV		

Aim of the WP

The goal of this WP is to derive an explainable model describing an AI system. The learned model will be compared in WP4-6 to the legal and moral principles formalised in WP3. For this reason, we will use Inductive Logic Programming (ILP) under the Answer Set Semantics (ASP) to learn a rule-based model of the AI system, relying on the deontic logic formalism for representing legal concepts as obligations, and extending it with moral concepts as preferences. In this way, it will be more straightforward to evaluate the compliance of the

learned model to the legal and moral value system. The objectives of this WP are then twofold: i) defining suitable moral deontic operators in ASP for LEGALIS's tasks; ii) designing an ILP task to learn moral deontic ASP rules.	
Tasks	
T2.1	**ASP extension to moral deontic logic (M1-M8: responsible UV; involved IIIA-CSIC, UB).** We will stem from existing literature focusing on the representation of deontic logical knowledge into the ASP formalism. In particular, LEGALIS aims not only at verifying compliance of AI systems with legal rules, but also with multiple ethical and moral values. While past works have focused mainly on representing legal knowledge bases in deontic ASP, e.g., obligations and permissions [Rob23], moral concepts as preferences and priorities are yet missing. On the other hand, they can be elegantly encoded in ASP, e.g., expressing preferences as weak constraints, subject to hard constraints imposed by the legal system, or as decision theoretic value functions. In this task, we will analyse the legal and moral expressiveness required by LEGALIS's tasks, in cooperation with the internal experts in the research team. We will then identify the relevant operators and concepts to represent them in moral deontic ASP.
T2.2	**ILP task design under moral, deontic ASP semantics (M9-M18: responsible UV, involved IIIA-CSIC, UB).** Given the moral, deontic extension of ASP, in this task we will learn a model of the AI system expressed in this formalism, using ILP. In particular, we will first identify the most suitable formalisation of ILP examples, encoding observations from the AI system. Typically, positive and negative examples are collected in ILP, encoding actual observations and forbidden situations, respectively. However, we will face two challenges: i) observations are inevitably uncertain (e.g., affected by imprecise behaviour of the agent or by sensors' uncertainty) and provide a partial description of the system; ii) negative examples are typically unavailable, since forbidden situations should never be observed. To overcome the first issue, we will adopt the most recent and efficient tools for ILP [Law20] to learn an ASP theory from noisy observations, where each positive example is associated with a weight defining its statistical reliability (defined in accordance with the AI stakeholders and the specific AI system), which can be interpreted as the relevance of the example with respect to the ASP theory. To address the second challenge, we will rely on the specific example structure of ILP, consisting of an included and excluded set, where the former encodes actual observations of the AI system, while the latter encodes unobserved facts. In this way, the ILP system will search for rules explaining not only the included set, but also the excluded set (e.g., negating unobserved facts). This will result in more detailed and specific rules for the underlying AI system. Indeed, the ILP system will search for rules . While this methodology does not

	semantically correspond to forbidding unobserved situations as negative examples would, it will allow to derive expressive and informative rules describing the main aspects of the AI system, corresponding to its nominal or standard behaviour. Moreover, ILP naturally provides not only the nominal rule system, but also the set of counterexamples, i.e., examples not explained by the theory (due, e.g., to the noise level). Counterexamples represent exceptions of the learned ASP theory, hence they give a valuable and informative input to counterfactual explanations in WP5, with the opportunity to provide a deeper insight into the learned model and the compliance of the AI system to the moral and legal values in specific situations, rather than only in the nominal case represented by learned rules. Finally, we will exploit the paradigm of ILP from ordered examples to define a priority rank between examples, according to stakeholder specifications. In this way, we will learn weak constraints describing preferences of the AI system in specific situations, approximating the moral and ethical behaviour of the system itself.
T2.3	**Evaluate and validate learned ILP model (M9-M20: responsible UV, involved UB, IIIA-CSIC).** In this task, we will evaluate the outcome of the above tasks T2.1-T2.2, supported by internal experts in the research team. This will serve as a fundamental validation step, before the final validation by external experts in the advisory board in WP6. Specifically, we will first assess the adequate expressiveness of the moral ASP deontic operators defined in T2.1, in the specific use cases of autonomous driving and caregiving robots. Then, we will assess the learned logical model (T2.2), evaluating the nominal rules induced via ILP and their relevance to our tasks. In this study, we will consider deontic and moral (i.e., preference) operators separately, in order to independently assess their importance for generating interpretable AI models. We will then analyse the counterexamples, using them as an input to counterfactual explanations (T5.1).

WP3	**Legal and ethical knowledge engineering**	**M1**	**M21**
Leader	**IIIA-CSIC**		
Aim of the WP • To build legal repositories that can be valuable for our two case studies and subsequently exploited by LLMs.			

<table>
<tr><td colspan="2">

- To investigate the RAG architecture (combining an LLM with external knowledge. from legal repositories) and fine-tuning with human feedback approach that yields best accuracy when retrieving applicable laws.
- To translate applicable laws into legal rules that can be subsequently verified.
- To encode value systems, containing the moral values and preference over them, which apply to our two case studies.

</td></tr>
<tr><td colspan="2">Tasks</td></tr>
<tr><td>T3.1</td><td>Building legal repositories (M1-M21: responsible IIIA-CSIC; involved UB, UV). To determine the laws that the AI under inspection is expected to comply with, a regulator must consider a vast number of hard laws (e.g., the EU AI Act) and soft laws (e.gstandards like the NIST AI RMF) [Cas24a,b,c], which will be available either as legal repositories or documents available in large legal repositories (such as B.O.E, Legifrance or Eurlex at the EU level). While many of them can be considered as raw data to be gathered, there is at least one publicly available Spanish repository containing structured data which is ready for computational handling (see the Spanish legal domain Language Model trained with legal corpora at https://github.com/PlanTL-GOB-ES/lm-legal-es). Another source to be considered is RoBERTalex, a transformer-based masked language model for the Spanish language, based on the RoBERTa model and pre-trained using a large Spanish Legal Domain Corpora, with a total of 8.9GB of text. As for the existing corpora containing European law, we will consider the LexFiles Corpus (19 billion tokens) https://github.com/coastalcph/lexlms, among other possibilities [Cha2023].
This task will focus on investigating, gathering, and pre-processing the available sources of legal information and organising them in “legal repositories” to be employed by the RAG architecture in task T3.2. Results from current projects such as ILENIA[40] will be leveraged.</td></tr>
<tr><td>T3.2</td><td>Develop a law AI companion for regulators (M6-M21: responsible IIIA-CSIC; involved UV, UB). Given the volume of legal documents and the need for an easy-to-use, preferably natural language-based querying interface for regulators, we propose to develop an AI assistant leveraging Large Language Models (LLMs). We will explore the use of Retrieval-Augmented Generation (RAG) for LLMs. This is currently the only known paradigm capable of improving the generation and instruction-following processes of LLMs by incorporating relevant domain information from available repositories or documents. Among the many available LLMs, we intend to begin our research from the Command R LLM, which, at the time of the writing, is the best (according to the widely-adopted "LLM Leaderboard" ranking)</td></tr>
</table>

[40] https://proyectoilenia.es/en/

	open-source model specifically designed for RAG. By doing so, we aim to address common challenges associated with LLMs, such as hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes, thus enhancing the accuracy and robustness of the resulting system [Gao+24]. We will investigate existing RAG architectures and evaluate/validate them with the assistance of legal experts to identify the most accurate one for retrieving relevant laws. In order to incorporate the feedback of such legal experts into our LLM without the drawbacks of classical supervised fine-tuning, we will explore the use of Direct Preference Optimization (DPO) [Raf24], a recent approach that has been shown to be more stable, performant, and computationally lightweight than Reinforcement Learning with Human Feedback (RLHF) [Chr17].
T3.3	**Legal rules' encoding (M9-M21: responsible IIIA-CSIC; involved UV, UB).** This task will investigate a method for translating applicable laws into legal rules. After applicable laws are retrieved for a given domain, we will encode them as legal rules in some formal language (e.g., deontic-event calculus [Has22, Cas24a]), so that they can be subsequently verified. Deontic logic allows representing concepts such as obligation or permissions, and it is used in the literature for model checking [Osm08] purposes. Alternatively, other formal rule-based formalisms equipped with different non-classical reasoning mechanisms, such as non-monotonic ones [AIJ80, Str+24] or argumentation based [Gar+04,Rah+09,Bar+18], will be explored in order to seek a compromise between expressiveness and computational effectiveness.
T3.4	**Value system encoding (M3-M21: responsible UB; involved IRI-CSIC, ICMAT-CSIC).** In the context of representing ethical knowledge, this process is akin to an engineering exercise. It commences with the establishment of a taxonomy of ethical objectives, such as the one used for self-driving cars in the study [Cab21] or the taxonomy designed for socially assistive robotics as explored in [Par21]. Subsequently, an interdisciplinary collaboration unfolds between applied ethics experts and AI specialists. Their joint effort aims to define ethical values and discern the preferences among them. In this framework, ethical values (such as safety or privacy) are construed as moral principles that serve as evaluative criteria for actions. Following this, the experts proceed to ascertain the relative preferences among various ethical values according to different policies (for example, prioritising safety over comfort in the driving scenario) [Per20, Nor22, Nor23a, Nor23b]. The outcome of this process is a value system, possibly expressed as a decision theoretic multi-attribute value function. Since the AI system under scrutiny is expected to align its behaviour with this value system, it will necessarily require domain-specific customization. Additionally, the selection of a reference value system for a given domain will involve the consideration of alternative

	value prioritizations and the analysis of the desirability of such alternative value systems.

WP4	Computing legal and ethical compliance	M3	M24
Leader	**UB**		
Aim of the WP ● to develop the computational tools for verifying whether an AI abides by applicable laws, as selected by legal experts, which have been encoded as rules (**legal compliance**) ● to develop the computational tools for verifying whether an AI behaves in alignment with a value system as specified by an applied ethics expert (**ethical compliance**) ● to internally validate the results output by our ethical and compliance algorithms with the aid of experts.			
Tasks			
T4.1	**Verifying legal compliance (M9-M24: responsible UV; involved ICMAT-CSIC, IIIA-CSIC).** To check legal compliance, we will consider two scenarios: either the AI input-output model is available for inspection (as a Deep Neural Network, DNN), or it is not, and we have to learn it from observations. In the first case, we will focus on a crucial weaknesses for DNN which is the vulnerability to adversarial input perturbations [Sze+13, Ami+23, Ins23].Specifically,, we will identify the input ranges and combinations which are invalid from a legal perspective. To this aim, we will leverage on the recent results of our research group [Mar+24], involving the efficient enumeration of unsafe (legally non compliant, in the context of LEGALIS)		

	inputs with probabilistic guarantees. Thanks to the probabilistic relaxation of the adversarial input problem, we will be able to extend the semantics of safety properties to be verified, to the rich semantics required by legal rules, e.g., deontic logics. This will allow us to identify specific behaviours of the AI system which are legally forbidden. In case the input-output model of the AI system is not available beforehand, we will first learn it from observations (T2.2), and then we will perform formal verification. While a DNN model could be learned, it is more convenient to directly learn an explainable logical model [Den16], e.g., in the ASP formalism. In this way, it will be possible to incorporate legal rules (defined by stakeholders) directly as constraints in the logical program, and formal verification of legal compliance will then consist in model checking [Rob23] or satisfiability checking at the reasoning stage, in order to assess whether the ASP program admits legal flows of execution.
T4.2	**Verifying ethical compliance (M3-M24: responsible UB; involved IIIA-CSIC, IRI-CSIC, ICMAT-CSIC).** Regarding ethical compliance, we will exploit the specification of ethical values and value systems produced by WP3 to compute the degree of *value alignment (ethical compliance)* [Nor23a, Nor23b, Gab20, Mon22, Bro21] of the AI under inspection. In particular, we will leverage our previous work [Mon22] to propose value alignment measures [Mon22]. Furthermore, we will use reinforcement learning techniques to train an AI that learns a policy exhibiting optimal value alignment. Thus, we can give regulators an ideal reference behaviour to compare, hence helping them understand how an ethical AI should behave. For that, we can leverage recent results by the research team, which allow us to train an AI to learn how to behave with optimal value alignment [Rod20, Rod21, Rod22, Rod23]. In particular, we will learn an explainable model of the AI system from observations (T2.2), expressing multiple value alignment in the ASP semantics. In this way, not only the AI model will be easily readable from experts and stakeholders, but we will be able to explicitly rank possible flows of execution of the AI system according to the ethical compliance [Den16]. In addition, the explainable model will incorporate legal constraints used in T4.1, resulting in a comprehensive representation for joint moral and legal verification. Besides, as a screening approach to both tasks 4.1 and 4.2, we shall adapt recent consortia work in relation to risk management (and compliance with the EU AI Act) of systems with AI-based components within the consortium [CAM24].
T4.3	**Evaluate and validate legal and ethical compliance (M9-M24: responsible UB; involved UV, IIIA-CSIC, ICMAT-CSIC, IRI-CSIC).** The purpose of this task will be to empirically evaluate, with the aid of the *internal experts* in the research team, the output produced by the algorithms developed by tasks T4.1 and

	T4.2 in our two case studies. The ultimate purpose of this evaluation will be to validate the output produced by our ethical and legal compliant algorithms. The results of the evaluation might lead to adjustments of the legal rules to verify (task T3.3), adjustments of the value system to verify (task T3.4), or refinements of the ethical and legal compliance algorithms (tasks T4.1 and T4.2) [Nor22, Nor23a, Cas24a, Cas24b].

WP5	**Explaining an AI and its legal and ethical compliance**	**M6**	**M24**
Leader	**ICMAT-CSIC**		
Aim of the WP To meet the stakeholders' right to an explanation, we will pursue two objectives: (i) building *counterfactual* explanations that help understand the behaviour of the AI under inspection, and (ii) explaining the results of assessing legal and ethical compliance.			
Tasks			
T5.1	**Building counterfactual explanations through causal reasoning (M6-M24: responsible ICMAT-CSIC; involved UV, IRI-CSIC, IIIA-CSIC).** The counterfactual explanation problem via causal reasoning will be first formulated in decision theoretic terms. Then, we shall provide methods that look for minimal input changes leading to a modification of a compliance recommendation (if positive to suggest the most relevant inputs in determining the recommendation; if negative, to suggest changes to reach positive compliance) or an alignment assessment (to suggest most sensitive change directions leading to improvement or worsening of alignment). Such analytic findings will be translated into natural language to facilitate understanding by non-sophisticated users and implemented and integrated with the other LEGALIS modules.		
T5.2	**Narratives and visual explanation to explain compliance (M9-M24: responsible UB; involved IIIA-CSIC, IRI-CSIC, ICMAT-CSIC).** This task will explain the results of the legal and ethical compliance assessment to end users. Since		

	we propose to use interpretable techniques to assess compliance, this goal becomes challenging from a human-computer interaction perspective when aiming at explaining the results comprehensively. Based on our contributions to this field [Kav23b, Zou22, Kav22], we will explore how to build *visual explanations* highlighting non-compliance with laws and misalignment with ethical values. Moreover, we will build on our previous work on chatbots [Kav24, Kav23a, Tel20] to consider the inclusion of a visualisation-oriented chatbot to make it easier, and more natural, for users to discover information from the explanations and provide them with effective analysis.

WP6	**Demonstrating and validating compliance**	**M1**	**M24**
Leader	**IIIA-CSIC**		
Aim of the WP ● To perform a thorough analysis of our two case studies that serves to identify legal and ethical issues that deserve compliance analysis. ● To build a software app that integrates our mechanisms for ethical and legal compliance (developed in WP4 and the explanatory mechanisms developed by WP5) and allows us to showcase the results of compliance analysis in our two use cases . ● To validate the results output by our ethical and compliance analysis with the aid of experts.			
Tasks			
T6.1	**Use cases' definition (M1-M6: responsible IIIA-CSIC; involved IRI-CSIC, ICMAT-CSIC, UB).** This task will focus on the thorough analysis of our two case studies to identify legal and ethical issues that deserve compliance analysis. On the one hand, we will analyse automated lane-changing decisions in heterogeneous traffic, which is one of the most challenging problems in relation to automated driving systems (ADS). For that, we count on the expertise of ICMAT-CSIC. On the other hand, we will consider an assistive robot that helps a caregiver to care for an elder needing help in basic activities like remembering to drink water, do cognitive or physical exercises, or keep social contact with their relatives. For that, we count on the expertise		

	of IRI-CSIC. In both cases, the research teams count on internal experts on: Applied Ethics (Begoña Román, UB), Law (Pompeu Casanovas, IIIA-CSIC), and Responsible AI (Maria Vanina Martínez, IIIA-CSIC). Furthermore, the use cases will be further evaluated and refined by the advisory board (M6). This task will feed the knowledge engineering tasks in WP3.
T6.2	**Building a software app that explains legal and ethical compliance (M12-M24: responsible IIIA-CSIC; involved UV, IRI-CSIC, ICMAT-CSIC, UB).** We will build a software application that will demonstrate the technological results of the project. We will equip the demonstrator with different AI models of autonomous cars, each one exhibiting different degrees of legal and ethical compliance. Along the same line, we will incorporate different AI models of a social assistive robot. The demonstrator will allow checking ethical and legal compliance using the algorithms developed by WP4. The demonstrator will allow a stakeholder to inquire about the behaviour of a car or a robot and receive counterfactual explanations produced by the resulting algorithm in task T5.1. It will also allow a user to perform a legal and ethical assessment. The demonstrator will report the results of the assessment using the results of task T5.2.
T6.3	**Validation with the aid of experts (M6-M24: responsible IIIA-CSIC; involved UV, IRI-CSIC, ICMAT-CSIC, UB).** We will perform two validation activities. First, we will validate the value systems and legal repositories to employ for ethical and legal compliance for our two use cases with the aid of the experts in law, ethics and responsible AI in the research team and in the advisory board. These are the results of tasks T3.1 and T3.4, which in turn are based on our case studies' analysis in task T6.1. Further ahead in the project, once our software tools and demonstrator are ready, we will showcase our demonstrator to our advisory board.

Figure 2 shows the Gantt diagram scheduling the work packages and tasks described above.

Figure 2. Gantt chart.

Deliverables. We will have one deliverable per work package per year. These are intended to deliver the results of the work carried out in that WP, usually as a set of articles (either accepted or being submitted for publications), and software code. Table 1 provides a summary of the deliverables. We note that WP1 does not have any corresponding deliverable because a final report will be submitted at the end of the project covering the work in this WP.

Deliverable #	Deliverable name	Work package	Leader	Due date

D2.1	Compiles the articles submitted during the first year within WP2 together with the definition of the ASP formalism extending moral deontic logic.	WP2	UV	M12
D2.2	Compiles the articles submitted during the second year within WP2 together with the software for ILP under the semantics of moral deontic ASP.	WP2	UV	M24
D3.1	*Legal and ethical knowledge engineering I:* Compiles the articles submitted during the first year within WP3, along with the first version of the Law AI companion software, and a report on the legal rules and value systems for the two use cases.	WP3	IIIA-CSIC	M12
D3.2	*Legal and ethical knowledge engineering II:* Compiles the articles submitted during the first and second year within WP3, along with a second version of the Law AI companion software, and a final report on the legal rules and value systems for the two use cases.	WP3	IIIA-CSIC	M24
D4.1	*Computing legal and ethical compliance I: Compiles the articles submitted during the first year within WP4, along with the first version of the compliance software and a preliminary report on the compliance of one of the two use cases.*	WP4	UB	M12
D4.2	*Computing legal and ethical compliance II:* Compiles the articles submitted during the first and second year within WP4, along with the final version of the compliance software, and the final report on the compliance of the two use cases.	WP4	UB	M24
D5.1	*Explaining an AI and its legal and ethical compliance I:* Compiles the articles submitted during the first year within WP5, along with the architectural design of the explanation software.	WP5	ICMAT-CSIC	M12
D5.2	*Explaining an AI and its legal and ethical compliance II:* Compiles the articles submitted during the second year within WP5, along with the implementation of the explanation software.	WP5	ICMAT-CSIC	M24

D6.1	*Demonstrating and validating compliance I:* First version of the demonstrator software	WP6	IIIA-CSIC	M12
D6.2	*Demonstrating and validating compliance II:* Second version of the demonstrator software and validation report	WP6	IIIA-CSIC	M24

Table 1. Deliverables list

Concerning the milestones, we have a milestone every six months to help detect and address potential risks, if any, and ensure the progress is timely. At M6, we will validate the design of the use cases. At M12, M18, and M24, we will validate the different versions of our developed models and mechanisms and their evaluation via the use cases. Note that advisory board members will be asked to help validate the milestones MS1 and MS3 at M6 and M18, respectively. Table 2 compiles the milestones to achieve in LEGALIS.

Milestone	Description	Work package	Due date	Verification
MS1	Value systems defined and legal repositories available for both use cases	WP2-WP6	M6	Assessed at the first consortium meeting, and by the advisory board
MS2	Initial results on LEGALIS ethical and legal compliance mechanisms, and their evaluation through the varied use cases	WP2-WP6	M12	Assessed at the second consortium meeting
MS3	Evolution of the ethical and legal compliance mechanisms, and the developed software tools and demonstrator	WP2-WP6	M18	Assessed at the third consortium meeting, and by the advisory board
MS4	Final results on LEGALIS' ethical and legal compliance mechanisms and software tools, and their evaluation through the proposed use cases	WP2-WP6	M24	Assessed at the consortium's closing meeting

Table 2. List of milestones

Risks and contingency plan

Table 3 below identifies potential risks and proposes mitigation measures to address them.

Risk	Description	WP	Proposed risk-mitigation measures
R1	Our analysis of ethical and legal issues in the use cases ignores relevant ethical and legal issues	WP6	We will show the result of our analysis to the advisory board (M6) to detect missing ethical and legal issues
R2	Lack of accuracy of retrieved applicable laws	WP3	Improve fine tuning with human feedback
R3	Ethical and legal compliance are not considered sufficiently valid by our internal experts	WP4	We will determine, in collaboration with our internal experts, whether we need to revise: the applicable laws retrieved by task 3.2 in collaboration with legal experts, the encoding of legal rules by task T3.3, the encoding of the value systems by task 3.4, or the compliance algorithms in tasks 4.1 and 4.2.
R4	Too computationally intensive explanation methods designed.	WP5	Reserve these as offline versions and design simplified (e.g., linear vs non-linear) versions for online use.
R5	The advisory board finds that the validation conducted by WP6 is not satisfactory enough	WP6	We leave six months after demonstrating results to the advisory board (M18) to perform further adjustments and refinements
R6	Due to the time limitation of the project, a delay in the results or insufficient quality of results may affect subsequent tasks.	WP2-WP6	Project coordination (see Task 1.1) includes frequent follow ups to continuously assess the advances in the different tasks in order to detect possible delays. Prompt detection is key for replanning and recovery

Table 3. Risks & contingency plan.

Advisory board

The project will have an advisory board, with the members below. All of them have already committed to be part of the advisory board (see accompanying support letters to this application).

- **Paula Boddington** authored the book "Towards a Code of Ethics for Artificial Intelligence" edited by Springer in 2017. Currently, she is Associate Professor of Philosophy and Healthcare at the School of Biomedical Sciences from the University of West London. Paula has a strong track record of interdisciplinary collaboration on research projects focusing on ethics and Artificial Intelligence (in particular with computer scientists in the Department of Computer

Science at Oxford University). Additionally, she is a member of the EU Atomium European Institute AI4People 2020 Committees.

- **Néstor Duch-Brown**, Scientific Officer and Team Leader at the Digital Economy Unit of the Joint Research Centre of the European Commission. He is an expert on Digital Economy and Policy-making.
- **Sarvapali Ramchurn**, Professor of Artificial Intelligence at the School of Electronics and Computer Science at the University of Southampton, CEO of Responsible AI UK, a £31m programme to develop and support an international ecosystem for responsible AI, Director of the UKRI Trustworthy Autonomous Systems Hub. AXA Research Award in 2018 for his work on Responsible AI.
- **Michael Luck**, Deputy Vice-Chancellor and Provost at University of Essex, founding Director of King's College London's Institute for Artificial Intelligence and Director of the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence, member of the Engineering & Physical Sciences Research Council (EPSRC) Strategic Advisory Team for Information and Communication Technologies (ICT).
- **Ugo Pagallo**, Law professor at University of Turin and European expert in law and technology (robotics and connected automated vehicles, CAV). The Japanese and Chinese editions of his Springer book on *The Laws of Robots* have been available since Spring 2018.
- **Louis de Koker**, Law professor at La Trobe University, attorney of the High Court of South Africa, fellow of the Society of Advanced Legal Studies, and World Bank expert on law, technology and development. He led the Australian Law and Policy Program of the Data to Decisions Cooperative Research Centre (CRC).

The plan is to meet with the advisory board twice. Once at M6 to explain our plans and the ongoing work, and get feedback that can be incorporated into our plans. And another time at M18 to help evaluate our work, with sufficient time (six months) to address the feedback received.

F/ Dissemination and exploitation plans

Dissemination initiatives towards the society & scientific community

We acknowledge the crucial importance of effectively disseminating and leveraging the project's outcomes to achieve the anticipated social and economic impact. Our focus lies in sharing the knowledge generated by the project with the research community and establishing relationships with various stakeholders to foster technology transfer opportunities. Within this framework, our effort will encompass:

- Coordinating internal and consortium communication efforts by devising a comprehensive communication plan encompassing the creation of communication materials such as news articles and press releases pertaining to project activities and outcomes. These messages will be disseminated through diverse channels and networks.
- Planning for scientific publications: Each sub-project leader will develop a publication plan. Since this is an interdisciplinary project, the research team will target publications in Artificial Intelligence, Robotics, Ethics, and Law. Thus, we will target reputable journals like the Artificial Intelligence Journal (AIJ), AI Magazine, Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS), International Journal of Social Robotics, ACM Transactions on Human-Robot Interaction, Minds and Machines, Ethics and Information Technology, Artificial Intelligence and Law, Technology in Society, the Journal of Business Ethics, Risk Analysis and Decision Support Systems. Additionally, relevant conferences such as IJCAI, AAAI, AAMAS, IROS, ICRA, HRI, AAAI/ACM Conference on AI, Ethics, and Society and Society for Risk Analysis have been identified for presentations and paper submissions.
- Facilitating stakeholder meetings: This involves identifying contacts, coordinating, and conducting meetings related to the execution, evaluation, and demonstration of use cases. These meetings aim to foster new collaborations and agreements.
- Organising an industry day: We will identify, coordinate, and host a workshop targeting companies in areas of impact and application to showcase project results, validate market opportunities, and explore the development of innovative products and services. These sessions will also seek to establish new collaborations and agreements.
- Intellectual Property (IP) Protection: Identifying promising results eligible for protection, conducting patent analysis, and implementing software registration procedures will be prioritised to safeguard valuable intellectual assets.

G/ Budget & financial sources

Most of the funding that we request will serve to hire personnel that spurs synergies between the research partners:

- **CSIC** plans to hire two full-time research engineers for two years each. One of them will support the research in WP3 (Legal and ethical knowledge engineering) and its connection with WP4, and the other one will support the research in WP5 (Explaining an AI and its legal and ethical compliance).
- **UV** plans to hire one full-time research engineer for two years to support the research in WP2 (Model learning from observation) and in task T4.1 (Verifying legal compliance).
- **UB** plans to hire one full-time research engineer for one year to support the research in tasks in WP3 (Legal and ethical knowledge engineering), WP4 (Computing legal and ethical compliance) and WP5 (Explaining an AI and its legal and ethical compliance).

All the contracted research engineers will collaboratively contribute to the development of the demonstrators in WP6. We have also allocated 7.000 EUR per partner to attend one international conference (3.000 EUR) and two European conferences (2.000 EUR) each. Finally, CSIC requests 10.000 EUR to fund a research stage of Dr. Mustafa Hashmi to assist in the building of legal repositories and extracting of legal rules from norms (task T3.1).

Financial Costs			
	CSIC	**UB**	**UV**
Personnel Costs	207.431,45	49.560	100.000
Research expenses other than personnel costs (equipment, licences etc.)	0	0	0
Travel, Meetings	17.000	7.000	7.000
TOTAL BUDGET PER PARTNER	**226.431,45**	**56.560**	**107.000**
REQUESTED BUDGET	**389.991,45**		

Table 4. Budget

H/ References

[AIJ80] AI Special Issue, 1980 Vol 13, 1-2 Nonmonotonic Reasoning.

[Ama21] Amantea, Ilaria Angela, Livio Robaldo, Emilio Sulis, Guido Boella, and Guido Governatori. "Semi-automated checking for regulatory compliance in e-Health." In *2021 ieee 25th international enterprise distributed object computing workshop (edocw)*, pp. 318-325. IEEE, 2021.

[Ami+23] Amir, G.; Corsi, D.; Yerushalmi, R.; Marzari, L.; Harel, D.; Farinelli, A.; and Katz, G. 2023. Verifying learning-based robotic navigation systems. In the 29th International Conference, TACAS 2023, 607–627. Springer.

[Amo+16] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." *arXiv preprint arXiv:1606.06565* (2016).

[Ash17] Ashley, Kevin D. 2017. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press.

[Bis+21] F. Bistaffa, C. Blum, J. Cerquides, A. Farinelli and J. A. Rodríguez-Aguilar, "A Computational Approach to Quantify the Benefits of Ridesharing for Policy Makers and Travellers," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 119-130, Jan. 2021, doi: 10.1109/TITS.2019.2954982.

[Bar+18] Baroni, Pietro ; Gabbay, Dov ; Giacomin, Massimilino & van der Torre, Leendert (eds.) (2018). Handbook of Formal Argumentation. London, England: College Publications.

[Ben12] Bench-Capon, Trevor, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier et al. "A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law." *Artificial Intelligence and Law* 20 (2012): 215-319.

[Bra+16] Brandt, Conitzer, Endriss, Lang, Procaccia. Handbook of computational social choice, Cambridge University Press, 2016.

[Bro21] Brown, D. S., Schneider, J., Dragan, A., & Niekum, S. (2021, July). Value alignment verification. In the International Conference on Machine Learning (pp. 1105-1115). PMLR.

[Bur21] Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245-317.

[Cab21] Caballero, W., Naveiro, R., Ríos Insua, D. (2021) Modeling Ethical and Operational Preferences in Automated Driving Systems. Decision Analysis 19(1):21-43.

[Cab23] Caballero, W., Rios Insua, D., Naveiro, R. (2023) Some statistical challenges in automated driving systems. Applied Stochastic Models in Business and Industry 39:5, pages 629-652.

[Cam24] Camacho, J.M., Couce-Vieira, A., Arroyo, D., Rios Insua, D. (2024) A Cybersecurity Risk Analysis Framework for Systems with Artificial Intelligence Components, arxiv 2401.01630.

[Can21] G. Canal, C. Torras, and G. Alenyà, "Are preferences useful for better assistance?: A Physically Assistive Robotics user study," ACM Transactions on Human-Robot Interaction, 10(4), pp. 1–19, 2021.

[Cas07] Casanovas, P., Noriega, P., Bourcier, D., & Galindo, F. (2007). *Trends in legal knowledge-The semantic web and the regulation of electronic social systems*. Florence: European Academic

[Cas16] Casanovas, Pompeu, Monica Palmirani, Silvio Peroni, Tom Van Engers, and Fabio Vitali. "Semantic web for the legal domain: the next step." *Semantic Web* 7, no. 3 (2016): 213-227.

[Cas17] Casanovas, P.; Rodríguez-Doncel, V.; González-Conejero, J. (2017). "The Role of Pragmatics in the Web of Data", in Capone, A.; Poggi F. (Eds.) *Pragmatics and Law. Practical and theoretical perspectives*. Dordrecht, Heidelberg: Springer, pp. pp 293-330.

[Cas24a] Casanovas, P., Hashmi, M., de Koker, I. & Lam, Ho-Pun (2024) "A methodological approach to legal governance validation", AICOL23, LNAI, Springer (in press)

[Cas24b] Casanovas, P. (2024). "Building a Smart Legal Ecosystem for Industry 5.0", W. Barfield, Y.H Weng, and U. Pagallo, Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction, Cambridge University Press, pp. 145-168.

[Cas24c] Casanovas, P., Hashmi, M., de Koker, L. & Pun-Lam, H. (2024). "Compliance, Regtech, and Smart Legal Ecosystems: A Methodology for Legal Governance Validation", in W. Barfield and U. Pagallo (eds.), Research Handbook on the Law of Artificial Intelligence, vol. II, Cheltenham (UK), Northhampton (MA): Edward Elgar Publ. (in press)

[Cha21] Chatila, Raja, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. "Trustworthy ai." Reflections on artificial intelligence for humanity (2021): 13-39.

[Chal23] Chalkidis, Ilias, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. "LeXFiles and LegalLAMA: Facilitating English multinational legal language model development." arXiv preprint arXiv:2305.07507 (2023).

[Chr17] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). "Deep reinforcement learning from human preferences." Advances in neural information processing systems, 30.

[Cro22] Cropper, A., & Dumančić, S. (2022). Inductive logic programming at 30: a new introduction. Journal of Artificial Intelligence Research, 74, 765-850.

[Den16] Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems, 77, 1-14.

[Dig+18] Dignum, Virginia, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova et al. "Ethics by design: Necessity or curse?." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 60-66. 2018.

[Dig19] Dignum, Virginia. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 1. Cham: Springer, 2019.

[Dom+16] Dominic, D.; Chhawri, S.; Eustice, R.M.; Ma, D.; Weimerskirch, A. Risk assessment for cooperative automated driving. In Proceedings of the 2nd ACM Workshop on Cyber-Physical Systems Security and Privacy, Vienna, Austria, 28 October 2016; ACM: New York, NY, USA, 2016; pp. 47–58

[EC23] European Commission. "Ethics guidelines for trustworthy AI." https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2023).

[EC24] European Commission. "EU Artificial Intelligence Act". https://artificialintelligenceact.eu/ (2024).

[Fac23] FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

[Flo19] Floridi, Luciano, and Josh Cowls (2019) "A unified framework of five principles for AI in society" (2019), *Harvard Data Science Review* 1.1: 1-15.

[Fos15] Villaronga, Eduard Fosch. "Legal and regulatory challenges for physical assistant robots." In *eChallenges e-2015 Conference*, pp. 1-8. IEEE, 2015.

[Fos21] Fosch-Villaronga, Eduard, and Tobias Mahler. "Cybersecurity, safety and robots: Strengthening the link between cybersecurity and safety in the context of care robots." *Computer law & security review* 41 (2021): 105528.

[Gab20] Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

[Gab+13] Gabrielli, Rinzivillo, Ronzano, Villatoro. "From tweets to semantic trajectories: mining anomalous urban mobility patterns". In: Proc. International Workshop on Citizen in Sensor Networks, 2013.

[Gall24] Gallego, V. (2024) Refined Direct Preference Optimization with Synthetic Data for Behavioral Alignment of LLMs, arxiv 2402.08005.

[Gar=04] García, A. J., & Simari, G. R. (2004). Defeasible logic programming: An argumentative approach. *Theory and practice of logic programming*, *4*(1-2), 95-138.

[Gao+24] Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* (2023).

[Gop+15] Gopalakrishnan, K.; Chitturi, M.V.; Prentkovskis, O. Smart and sustainable transport: Short review of the special issue. Taylor & Francis: Abingdon, UK, 2015.

[Gov16] Governatori, Guido, Mustafa Hashmi, Ho-Pun Lam, Serena Villata, and Monica Palmirani. "Semantic business process regulatory compliance checking using LegalRuleML." In *European Knowledge Acquisition Workshop*, pp. 746-761. Cham: Springer International Publishing, 2016.

[Gov17] Governatori, Guido. "A Short Introduction to the Regorous Compliance by Design Methodology." In *TERECOM@ JURIX*, pp. 7-13. 2017.

[Gio13] Giordano, L., Martelli, A., & Dupré, D. T. (2013, June). Temporal deontic action logic for the verification of compliance to norms in ASP. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (pp. 53-62).

[Har16] Harmon, Paul. 2016. "The State of Business Process Management - A BPTrends Report." http://www.bptrends.com/bpt/wp-content/uploads/2015-BPT-Survey-Report.pdf.

[Has15] Hashmi, M. . "A methodology for extracting legal norms from regulatory documents." In *2015 IEEE 19th International Enterprise Distributed Object Computing Workshop*, pp. 41-50. IEEE, 2015.

[Has18a] Hashmi, M., Governatori, G., Lam, B. & Wynn, M.T. (2018). "Are we done with business process compliance: state of the art and challenges ahead", *Knowledge and Information Systems* (Vol.57 n.1, pp. 59-133).

[Has18b] Hashmi, Mustafa, Pompeu Casanovas, and Louis De Koker. "Legal compliance through design: preliminary results of a literature survey." *TERECOM2018@ JURIX, Technologies for Regulatory Compliance http://ceur-ws. org* 2309 (2018): 06.

[Has22] Hashmi, M. "On Modelling Process Aspects With Deontic Event-Calculus" , *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)* 13, no. 1 (2022): 1-19.

[Ins90] Insua, D.R. (1990) Sensitivity Analysis in Multiobjective Decision Making, Springer.

[Ins23] Insua, D.R., Naveiro, R., Gallego, V., & Poulos, J. (2023). Adversarial Machine Learning: Bayesian Perspectives. Journal of the American Statistical Association, 118(543), 2195–2206.

[Ins22] Insua, D.R, Caballero, W., Naveiro, R. (2022)Managing Driving Modes in Automated Driving Systems, Transportation Science, 56, 1259-1278.

[Kar22] Karimi, A., Barthe, G., Schölkopf, B., Valera, I. (2022) A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. ACM Comput. Surv. 55, 5.

[Kau22] Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. ACM computing surveys (CSUR), 55(2), 1-38.

[Kav24] Kavaz, E., Wright, F., Nofre, M., Puig, A., Rodríguez, I., Taulé, M. Introducing the Multidisciplinary Design of a Visualisation-Oriented Natural Language Interface, Accepted CEDI-2024.

[Kav23a] Kavaz, E., Puig, A., Rodríguez I. 2023. Chatbot-Based Natural Language Interfaces for Data Visualisation: A Scoping Review. Applied Sciences. https://doi.org/10.3390/app13127025

[Kav23b] Kavaz, E., Rodríguez, I., Puig, A., & Vives, E. (2023). A Conversational Data Visualisation Platform for Hierarchical Multivariate Data, EuroVis 2023.

[Kav22] Kavaz, E., Puig, A., Rodríguez I. Chacón, R., De-La-Paz, D., Torralba A., Nofre, M., Taulé, M. 2022. Visualisation of hierarchical multivariate data: Categorisation and case study on hate speech. Information Visualization. 22-1, pp.31-51. https://doi.org/10.1177/14738716221120509

[Lam09] Lam, Ho-Pun, and Guido Governatori. "The making of SPINdle." In *Rule Interchange and Applications: International Symposium, RuleML 2009, Las Vegas, Nevada, USA, November 5-7, 2009. Proceedings 3*, pp. 315-322. Springer Berlin Heidelberg, 2009.

[Lam18] Lam, Ho-Pun, Mustafa Hashmi, and Akhil Kumar. "Towards a Formal Framework for Partial Compliance of Business Processes." In *International Workshop on AI Approaches to the Complexity of Legal Systems*, pp. 90-105. Cham: Springer International Publishing, 2018.

[Lam19] Lam, Ho-Pun, and Mustafa Hashmi. "Enabling reasoning with LegalRuleML." *Theory and Practice of Logic Programming* 19, no. 1 (2019): 1-26.

[Law20] Law, M., Russo, A., Bertino, E., Broda, K., & Lobo, J. (2020, April). Fastlas: Scalable inductive logic programming incorporating domain-specific optimisation criteria. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 03, pp. 2877-2885).

[Lei+17] Leike, Jan, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. "AI safety gridworlds." *arXiv preprint arXiv:1711.09883* (2017).

[Lif19] Lifschitz, V. (2019). Answer set programming (Vol. 3). Heidelberg: Springer.

[Lov24] T. Love, A. Andriella, and G. Alenyà, "Towards explainable proactive robot interactions for groups of people in unstructured environments," in ACM/IEEE International Conference on Human-Robot Interaction, 2024, pp. 697–701.

[Mad+20] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable Reinforcement Learning through a Causal Lens. Proceedings of the AAAI Conference on Artificial Intelligence, 34(03):2493–2500, April 2020.

[Mad+22] Madiega, T., Van De Pol, Anne Louise. Artificial intelligence act and regulatory sandboxes. European Parliamentary Research Service. PE 733.544 – June 2022

[Mar09] Martens, D., Baesens, B. B., & Van Gestel, T. (2008). Decompositional rule extraction from support vector machines by active learning. IEEE Transactions on Knowledge and Data Engineering, 21(2), 178-191.

[Mar+24] Marzari, Luca, Davide Corsi, Enrico Marchesini, Alessandro Farinelli, and Ferdinando Cicalese. "Enumerating safe regions in deep neural networks with provable probabilistic guarantees." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 19, pp. 21387-21394. 2024.

[Mcg23] McGreal, Paul E. "Corporate compliance survey." *The Business Lawyer* 77, no. 2 (2023): 475-496.

[Mei24] Meixide, C.G., Ríos Insua, D. (2024) Generative invariance: causal extrapolation without exogeneity, arxiv 2402.15502.

[Mel+21] Meli, et al., Inductive learning of answer set programs for autonomous surgical task planning: Application to a training task for surgeons, Machine Learning (2021).

[Mel+24] Meli, D., Castellini, A., & Farinelli, A. (2024). Learning Logic Specifications for Policy Guidance in POMDPs: an Inductive Logic Programming Approach. Journal of Artificial Intelligence Research, 79, 725-776.

[Mil19] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.

[Mon22] Montes, N., & Sierra, C. (2022). Synthesis and properties of optimally value-aligned normative systems. Journal of Artificial Intelligence Research, 74, 1739-1774.

[Mor22] Morgan AA, Abdi J, Syed MAQ, Kohen GE, Barlow P, Vizcaychipi MP. Robots in Healthcare: a Scoping Review. Curr Robot Rep. 2022;3(4):271-280.

[Mug91] Muggleton, S. Inductive logic programming. New Generation Computing (1991).

[Nav22] Naveiro, Roi and Caballero, William N. and Ríos Insua, David, Adversarial Risk Analysis for Heterogeneous Traffic Management. Available at SSRN: https://ssrn.com/abstract=4365987.

[Nay22] Nay, J.J. "Law informs code: A legal informatics approach to aligning artificial intelligence with humans." *Nw. J. Tech. & Intell. Prop.* 20 (2022): 309-393.

[Nay24] Nay JJ, Karamardian D, Lawsky SB, TaoW, Bhat M, Jain R, Lee AT, Choi JH, Kasai J. 2024 Large language models as tax attorneys: a case study in legal capabilities emergence. Phil. Trans. R. Soc. A 382: 20230159. https://doi.org/10.1098/rsta.2023.0159

[Nor23a] Noriega, P., Verhagen, H., Padget, J., d'Inverno, M. (2023). Addressing the Value Alignment Problem Through Online Institutions. In: Fornara, N., Cheriyan, J., Mertzani, A. (eds) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI. COINE 2023. Lecture Notes in Computer Science(), vol 14002. Springer, Cham. https://doi.org/10.1007/978-3-031-49133-7_5

[Nor23b] Noriega, Pablo, Harko Verhagen, and Julian Padget. "Design Heuristics for Ethical Online Institutions." Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV: International Workshop, COINE 2022, Virtual Event, May 9, 2022, Revised Selected Papers. Vol. 13549. Springer Nature, 2022.

[Oec23] OECD. Regulatory Sandboxes in Artificial Intelligence. OECD Digital Economy Papers. July 2023 N° 356.

[Oli23] de Oliveira Rodrigues, Cleyton Mario, Frederico Luiz Gonçalves de Freitas, Emanoel Francisco Spósito Barreiros, Ryan Ribeiro de Azevedo, and Adauto Trigueiro de Almeida Filho. "Legal ontologies over time: A systematic mapping study." *Expert Systems with Applications* 130 (2019): 12-30.

[Osm08] Osman, Nardine Zoulfikar. "Runtime verification of deontic and trust models in multiagent interactions." (2008).

[Par21] Pareto Boada, J., Román, B., and Torras, C. 2021. The ethical issues of social assistive robotics: A critical literature review. Technology in Society, 67: 101726.

[Par20] Pareto, J., Román, B.. 2020. Reforming Public Administration: The Codes Of Ethics In City Councils. Ramon Llull Journal of Applied Ethics.

[Per20] Antoni Perello-Moragues, & Pablo Noriega (2020). Using Agent-Based Simulation to Understand the Role of Values in Policy-Making. Harko Verhagen, Melania Borit, Giangiacomo Bravo, & Nanda Wijermans (Eds.), Advances in Social Simulation (pp. 355--369). Springer International Publishing.

[Per21] Antoni Perello-Moragues, Pablo Noriega, Lucia Alexandra Popartan, & Manel Poch (2021). On Three Ethical Aspects Involved in Using Agent-Based Social Simulation for Policy-Making. Petra Ahrweiler, & Martin Neumann (Eds.), Advances in Social Simulation (pp. 415--427). Springer International Publishing.

[Pob19] Poblet, M., Casanovas, P. and Rodríguez-Doncel, V., 2019. Linked Democracy: Foundations, tools, and applications. Cham: Springer Nature.

[Pea+16] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. Causal inference in statistics: a primer. Wiley, Chichester, West Sussex, 2016.

[Pet+15] Petit, J.; Shladover, S.E. Potential cyberattacks on automated vehicles. IEEE Trans. Intell. Transp. Syst. 2015, 16, 546–556.

[Raf24] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). "Direct preference optimization: Your language model is secretly a reward model." Advances in Neural Information Processing Systems, 36.

[Rah+09] Rahwan, I., & Simari, G. R. (Eds.). (2009). Argumentation in artificial intelligence. Dordrecht: Springer.

[Ras+23] Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. 2022. Explainable Deep Learning: A Field Guide for the Uninitiated. J. Artif. Int. Res. 73 (May 2022). https://doi.org/10.1613/jair.1.13200.

[Rob23] Robaldo, L., Batsakis, S., Calegari, R., Calimeri, F., Fujita, M., Governatori, G., ... & Zangari, J. (2023). Compliance checking on first-order knowledge with conflicting and compensatory norms: a comparison among currently available technologies. Artificial Intelligence and Law, 1-51.

[RodD16] Rodriguez-Doncel, V., Santos, C., Casanovas, P. and Gómez-Pérez, A., 2016. "Legal aspects of linked data–The European framework". *Computer law & security review*, *32*(6), pp.799-813.

[RodD21] Rodríguez-Doncel, Víctor, Monica Palmirani, Michał Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor (eds.). *AI Approaches to the Complexity of Legal Systems XI-XII*. LNCS 13048, Springer International Publishing, .

[Rod20] Rodriguez-Soto, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2020, May). A structural solution to sequential moral dilemmas. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (pp. 1152-1160).

[Rod21] Rodriguez-Soto, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2021). Multi-Objective Reinforcement Learning for Designing Ethical Environments. In IJCAI (Vol. 21, pp. 545-551).

[Rod22] Rodriguez-Soto, M., Serramia, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2022). Instilling moral value alignment by means of multi-objective reinforcement learning. Ethics and Information Technology, 24(1), 9.

[Rod23] Rodriguez-Soto, M., Lopez-Sanchez, M. & Rodriguez-Aguilar, J.A. Multi-objective reinforcement learning for designing ethical multi-agent environments. Neural Computing & Applications (2023). https://doi.org/10.1007/s00521-023-08898-y

[Rot23] Rotolo, Antonino, and Giovanni Sartor. "Argumentation and explanation in the law." *Frontiers in Artificial Intelligence* 6 (2023): 1130559.

[Roy15] Royo-Bordonada, M.A., Román-Maestre, B. 2015. Towards public health ethics Public Health Reviews. Annual Review of Public Health. WOS https://doi.org/10.1186/s40985-015-0005-0

[Ser23] Serramià, M., Lopez-Sanchez, M., Rodríguez-Soto, M., Bistafa, F., Rodriguez-Aguilar, J.A. Boddington, P., Ansotegui, C. & Wooldridge, M. (2023). Encoding ethics to compute value-aligned norms. Minds and Machines. (doi: https://doi.org/10.1007/s11023-023-09649-7).

[Slo05] Steven Sloman. Causal Models: How People Think about the World and Its Alternatives. Oxford University PressNew York, 1 edition, August 2005.

[Str+24] Strasser, Christian and G. Aldo Antonelli, "Non-monotonic Logic", The Stanford Encyclopedia of Philosophy (Summer 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/sum2024/entries/logic-nonmonotonic.

[Sze+13] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[Tel20] Tellols, D.; Lopez-Sanchez, M.; Rodríguez-Santiago, I.; Almajano, P.; Puig, A.. 2020. Enhancing sentient embodied conversational agents with machine learning. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2019.11.035

[Ver+23] Veronese, Celeste, Daniele Meli, Filippo Bistaffa, Manel Rodríguez-Soto, Alessandro Farinelli, and Juan A. Rodríguez-Aguilar. "Inductive Logic Programming For Transparent Alignment With Multiple Moral Values." In CEUR WORKSHOP PROCEEDINGS, pp. 84-88. 2023.

[Vin+20] Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. "The role of artificial intelligence in achieving the Sustainable Development Goals." Nature communications 11, no. 1 (2020): 1-10.

[Von51] Von Wright, G. H. (1951). Deontic logic. Mind, 60(237), 1-15.

[Yue+23] Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. Explainable Reinforcement Learning via a Causal World Model, May 2023. arXiv:2305.02749.

[Yue23] Yue, Shengbin, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou et al. "Disc-lawllm: Fine-tuning large language models for intelligent legal services." *arXiv preprint arXiv:2309.11325* (2023).

[Zou22] Zoumpekas, T.; Salamó, M.; Puig, A. 2022. Rethinking Design and Evaluation of 3D Point Cloud Segmentation Models. Remote Sensing. https://doi.org/10.3390/rs14236049