Beyond the Digital Judge: Legal Reasoning in Compliance Checking and Compliance Choices

Marcello Ceci$^{[0000-0003-3800-0906]}$ and Domenico Bianculli$^{[0000-0002-4854-685X]}$

University of Luxembourg, Luxembourg
marcello.ceci@uni.lu, domenico.bianculli@uni.lu

Abstract. This paper investigates the practical reasoning involved in compliance-related decisions, distinguishing between two scenarios where a state of affairs is evaluated in the light of applicable norms: *ex post* compliance checking and *ex ante* compliance choices. While the literature in legal reasoning representation is exclusively focused on compliance checking scenarios, i.e., simulating a *digital judge*, different factors seem to play a role in the *inner deliberation* of compliance choices. In this paper we investigate how human agents are influenced in their compliance choices by their own value ranking and risk assessment and how, in turn, the choice affects their preference among alternative interpretations of the law. We contend that contributions from the literature such as the value-based argumentation framework, while focused on *ex post* judgments, may be able to provide a comprehensive framework for *ex ante* compliance decisions. The main goal of the research behind this work is to achieve a comprehensive representation of legal reasoning that can be used as a reference for the explanation of automated compliance decisions.

Keywords: AI and Law · Compliance · Digital Justice · Legal Reasoning · Human Decision Making · Value-based Argumentation.

1 Introduction

The ever-advancing process of automation is stepping into the area of decision making, including legal decisions affecting the personal freedom of individuals. As technological advancements go, it comes with its own dangers and inevitable learning curves. For example, a decision support software[1], used for decades in courts to automate the parole system, has been criticized on multiple occasions for racial bias [1, 5]. While the bias in this case is to be attributed to historical data, the risk of biased decisions is further increased for generative models, where the factors leading to a decision are even more opaque. Moreover, while important, the decision on parole concerns the execution of the penalty, and as such is not the most important decision in a legal process: other, more important decisions are likely to be automated in the future, such as the judgment of being guilty of breaching the law.

[1] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support software used in some U.S. courts to assess the likelihood of a defendant becoming a recidivist.

The prospect of Artificial Intelligence being applied not only as *digital judges* in courts of law, but also as *digital agents* in decisional processes that are subject to regulations[2] poses a critical question regarding our capability to formally represent and thus evaluate those types of decisions. Thus, there is a need to *represent the reasoning of digital agents faced with a compliance-related decision.*

Let us consider two cases, regarding natural and legal persons respectively, inspired by real-life events but anonymized due to privacy concerns:

- *Euthanasia:* a doctor assists consensual patients in performing euthanasia in a country where the practice constitutes a crime *unless its sole purpose is to reduce pain*. The doctor is arrested and tried several times: at first acquitted, then eventually convicted for second-degree murder.
- *Anti-money laundering (AML):* a cryptocurrency firm willingly facilitates transactions from sanctioned groups such as ransomware hackers, and is eventually fined for a hefty sum by the competent AML public authority.

As we will see, the research in AI and Law has done considerable advancements in the representation of the reasoning involved in those judges' decisions, but did not pay as much attention to the representation of the reasoning that may have led the doctor (in the euthanasia case) and the crypto firm (in the AML one) to act the way they did. However, scenarios relegated to science fiction just a few years ago, such as a digital doctor autonomously performing end-of-life choices in an automated healthcare institution or a nonhuman entity autonomously facilitating transactions, are not hard to imagine nowadays, and are likely to become reality in the near future [22]. A formalization and representation of these reasoning processes is paramount to achieve a framework against which compliance-related decisions of digital agents can be coherently explained and thus evaluated.

In this paper we investigate compliance-related decisions, i.e., decisions between alternative actions in the presence of normative restrictions. We do so from the perspective of legal knowledge representation and a focus on practical reasoning[3], with the purpose of examining the capability of the state-of-the-art to represent such reasoning for software applications and automation. In other words, we aim to understand whether our representation of the *digital judge* is on par with that of the *digital agent*. The reasoning involved in compliance choices is in fact different from that involved in other legal reasoning such as the one aimed at resolving conflicts between norms, which rely on rule priorities[4]: compliance decisions are influenced by several factors, including the *context* of the action and the *value ranking* of the agent.

To investigate the differences in legal reasoning, we introduce the two scenarios of *compliance checking* (an *ex post* judgment on a case) and *compliance choice* (an *ex ante* decision on which action to take in a situation). The state-of-the-art in checking the

[2] Many decisions are already automated, e.g., granting a loan in banking services. Future scenarios go way beyond that: Gervais and Nay [22] envision a *legal singularity* where nonhuman entities that are not directed by humans may enter the legal system as a new *species* of legal subjects, leading to an *interspecific* legal system.

[3] Practical reasoning is an approach to action selection proposed by Raz [42].

[4] Examples of priorities are *lex superior*, *lex posterior*, *lex specialis*.

compliance of an autonomous system with regulations, focusing on *ex post* approaches, tends to neglect the fact that breaking the law is sometimes desirable [7]. On the contrary, an autonomous system unable to understand the factors that influence a compliance choice will treat all breaches of the law in the same way, a dangerous scenario especially as the world become less democratic, with freedom of expression, deliberation, rule of law and elections in net decline in the last decade [26]. An automated compliance approach aiming for blind application of the law, neglecting its principles and the values shared by its subjects, has the potential to trigger a feedback loop [40] towards thoughtless, conformance-driven dystopias, whereas democratic systems aim to elicit rule-following on the basis of shared legal principles and values.

The rest of the paper is structured as follows. We begin our investigation into the representation of compliance decisions by presenting (Sect. 2) the recent advancements in the representation of legal reasoning. We then introduce (Sect. 3) the two scenarios of *ex post* and *ex ante* compliance decisions, successively focusing on the *inner deliberation* process of *ex ante* compliance choices and its relation with the preference among alternative interpretations of the law. We further show (Sect. 4) how additional factors seem to play a role in compliance choices, in a defeasible meta-argumentative process related to the (expressed) preference among alternative interpretations of a norm. We finally show (Sect. 5) how such a process can be represented via an argumentation framework, to support practical reasoning.

2 Background: the Representation of Legal Reasoning

In this section, we present the most important contributions in the representation of legal reasoning from the fields of AI and Law and Requirements Engineering (RE).

Models for Representing Legal Concepts. The field of AI and Law provides approaches for legal compliance checking based on modeling deontic norms, representing peculiarities of legal knowledge such as norm defeasibility or multiple interpretations [20]. In the field of RE, models such as the taxonomy of semantic metadata of legal provisions [46] or the approach for deriving business processes from the law [15] are attempts to tame the complexity of the legal domain with task-automation objectives in mind. However, those models use a high-level, often generic, representation of the provisions and exhibit a weak relationship to the real-life states of affairs [29].

Rule Languages for Representing Norms. In AI and Law, semantic markup languages (e.g., LegalRuleML) or open formats [20], and representations of the norms complexities [2] have been widely investigated. Complex normative structures have been studied also in RE, to support the precise definition of software requirements. While earlier attempts at rule-based formalizations lack consideration for legal peculiarities [13], more recent approaches such as Legal GRL by Ghavanati et al. [23] and Nòmos by Ingolfo et al. [27] account for aspects such as Hohfeldian relations and alternative representations of legal norms. Legal GRL represents legal requirements in the context of goal-based reasoning, thus integrating aspects of practical reasoning; however, while successfully representing the elements of legal reasoning applied to software engineering, the approach lacks consideration of the dynamics of it, as shown by Ghavanati et al.'s [23] assumption that "when a legal goal is of type *Obligation*, it must be satisfied completely". As we will see in Sect. 3.2, this is not always the case.

Legal Reasoning. While the field of RE does not focus on the automated application of legal requirements through reasoning, contributions on the topic from AI and Law are many and diverse. The most comprehensive solution to representing legal reasoning while accounting for aspects such as alternative representations is considered to be the use of argumentation [9, 10, 31], whose formal study has played an important role within Artificial Intelligence for considerable time [35]. Dung's abstract argumentation framework [18] provided a standard approach for detecting conflicts of arguments, with a huge impact on AI and Law [6]. Further research by Modgil and Prakken [37] led to the development of an extended argumentation framework, allowing the resolution of such conflicts via a meta-argumentation layer prioritizing arguments *pro* or *con* a certain conclusion according to preferences. Walton et al. [49] introduced the concept of argumentation schemes, forms of argument that capture stereotypical patterns of human reasoning. These schemes allow for the defeasibility of arguments via critical questions, enabling the representation of a meta-argumentation layer where a certain value ranking may determine the preference among conflicting arguments. Normative reasoning has been also captured in constrained I/O logics [33], and recent efforts are trying to combine this approach with formal argumentation [8]. The above presented contributions however focus on a judgment perspective (*ex post* analysis), neglecting the agent's perspective, as we will see next.

3 Towards a General Legal Reasoning Framework

3.1 Legal reasoning in *Ex post* Compliance Checking vs. *Ex ante* Compliance Choices

The literature in legal reasoning representation deals with ***ex post* compliance decisions**, i.e., the application of the law to a *case*[5]. These are the decisions taken by judges in judgments (constituting the body of norms called *case law*) and by compliance officers in internal audits. *Ex post* decisions lead to a *judgment of action*: we will therefore call this activity **compliance checking**. Since these evaluations concern past facts and events whose truth is approximated according to the burden of proof, the legal reasoning performed therein does not involve uncertain information.

However, since the purpose of compliance checking is to evaluate a past state of affairs from a strictly legal viewpoint (i.e., to determine if an action or state of affairs is in breach of applicable norms), with the exception of criminal trials[6] and theoretical work

[5] *Case* here is meant as defined by Čyras and Lachmayer [17], as opposed to *situation*. Situations concern the future and are related with ex-ante analysis, whereas cases are related with ex-post, being concerned with the past; this implies that certain judgments (e.g., preliminary rulings) which are not taken *ex post* are not considered in the present work.

[6] In the euthanasia case, the goals of the doctor are relevant to the judgment: the debate in court would revolve around whether the act had the *intent to relieve pain* as opposed to the *intent to cause death* (see Section 3.3). On the other hand, in the AML case, with AML being in the subdomain of corporate criminal liability, little to no importance is given to goals; in the example, the fact that illegal transactions were facilitated would alone justify the penalty.

on argumentation frameworks (see Sect. 5), current conceptualizations of compliance are not concerned with understanding why the illegal action was taken nonetheless. We have no clue as to whether the action was due to a mistake, a wrong interpretation, an unlucky bet, or a deliberate choice. Furthermore, we do not know what a rational person would do if faced by the same choice. This knowledge lies in the realm of **compliance choices**, i.e., ***ex ante* compliance decisions** on which action to take in a specific *situation* [17]. Since those decisions have to be taken on the basis of incomplete information and uncertain outcomes, they imply practical reasoning, following one of two approaches: either an *actualist approach*, i.e., they can try to *actually* obtain the best reward, or a *possibilist approach*, i.e., they can try to maximize the potential reward, however unlikely that might be[7]. While, under a possibilist approach, an agent will try to comply to the maximum possible extent with normative restrictions, human agents are rather prone to adopt the actualist approach, which implies a different way to factor in normative restrictions, as we will see in the next subsection.

3.2 The Inner Deliberation

Current approaches to compliance tend to assume that people's (and companies') behavior is perfectly law-abiding and thus provide a vision of the world according to which the agent acts exclusively within the boundaries of the law. For example, in the field of Requirements Engineering (RE), legal requirements are specified as if all norms were equally sufficient in influencing behavior. In the RE literature there is no mention of norms that can be ignored or breached, and how to determine them and translate that evaluation into software requirements. In autonomous systems, prohibited actions tend to be downright unavailable [7]. But is that truly how our world works? Do we comply to all norms that are applicable to our situation? For example, do we read and understand all the terms and conditions before using a software? Studies suggest otherwise [32, 47]. In fact, agents use a form of utility function to make decisions on whether to act in compliance or in breach of applicable norms, and normally choose the action which has the most favorable consequences *for themselves*.

Bench-Capon and Modgil [7] state that in some cases we know that a rule applies but we still do not conform. They go further, stating that unreflecting adherence to norms (i.e., blind rule conformance) is not what we expect from a genuinely moral reasoner: *rules are made to be broken*. This is indeed a step towards an *actualist* approach to compliance choices, where a deliberate breach of a norm is possible. In fact, when faced with a choice of action in the face of normative restrictions, we engage in an **inner deliberation**, an *actualist* dialogue juxtaposing conflicting arguments about whether to comply or not. Differently from compliance dialogues (discussed in the context of multi agent systems by Rotolo [43]) and deliberation dialogues (described by Atkinson and Bench-Capon [4]), the dialogue of inner deliberation does not involve multiple agents (or audiences) with different goals (or value rankings), but rather involves different potentially conflicting factors arguing for alternative choices of action.

When dealing with how humans act in the face of normative restrictions, we are in the domain of decisions in the face of uncertain outcomes. This correspond to the

[7] In value learning the two approaches are called *on-policy* and *off-policy* — see Christian [16].

domain of practical reasoning (or game theory when involving other agents), where we want to understand the possible implications in each choice, factoring the legal consequences (mainly, the risk of a penalty) with all other factors involved in decision making. The basic example to express the reward system in game theory considers that the cost of riding a bus without a ticket corresponds to the penalty for the violation multiplied by the enforcement probability: for example, if the fine for being caught without a valid ticket is 10€ and the chance of being caught is 10%, then riding ticketless is a rational choice if the price of a ticket exceeds 1€. In the world of legal compliance these considerations form part of *risk management*[8].

There is, however, more than risk management in a compliance choice, as greed is not the only reason for an agent to deliberately breach a norm. A breach may in fact be *necessary* in the situations when compliance to all applicable rules at once is impossible (e.g., because of a conflict of norms or jurisdictions). There are further cases, where the breach of a law is not simply a mistake, a bet, or an inevitable choice. For example, it can be a way to improve security: de Mattos et al. [34] report how 25% of surveyed papers consider breaches as *positive* (e.g., in improving work conditions and safety). In the euthanasia case, the doctor chose to assist people in causing their own death with the intention to do good (relieving pain, fulfilling a man's will): in fact, he never received any economic compensation for his services. After being prosecuted and acquitted several times in situations where he did not materially administer the lethal injection, he came to the point of administering it himself and publicly confessing his deed, practically requesting to be prosecuted, with the aim of seeing the lawfulness of any such acts sanctioned in a court of law. This configures as a blatant breach of a norm, i.e., a breach not caused by a mistake or a material impossibility to comply, and not done in secret in the hope to avoid legal scrutiny.

We have seen how laws can be breached intentionally, sometimes even deliberately. Before investigating further (in Sect. 4) the factors that lead a subject to choose to act in breach of a norm, we will now investigate the qualification of breaches: are they always acts of rebellion to the authority, or may they be something else? It turns out that behaviors judged as breaches are often not seen by the agents as deliberate violations, but rather as behaviors that are lawful, *under the agent's own interpretation of the norm*.

3.3 Arguing about Alternative Interpretations

There is always a limit to a norm's applicability: it cannot be overstretched to have effects that go against the principles behind it. These principles are often implicit, but can manifest themselves in the preamble of the law or through the so-called *final rules* (e.g., as theorized in Italy's legal system as *norme di chiusura*), whose function is to solve conflicts or gaps in the law. For example, in work safety law, even in the absence of specific norms, the foreman is still required to adopt caution and diligence in order to protect the physical integrity of workers[9]. Similar rules prevent legal norms to counteract the very principles they are meant to foster, e.g., if complying with a norm would

[8] Compliance risk [19] is a subcategory of operational risk.

[9] Art. 2087 of the Italian Civil Code.

harm the interest of third parties[10]. These norms constitute a particular kind of legal exceptions, whose effect is represented in AI and Law as rule *defeasibility* [10] rather than rule violation: from a legal-theoretical point of view, the agent is not breaching the norm but it is rather the norm's coercitive power to be defeated by an exception derived from legal principles. In case-based reasoning, an influential contribution by Berman and Hafner [9] argues that, in absence of precedent, the decision should be made according to which social purposes would be promoted by deciding for either part, choosing whichever would better serve the prevalent social values.

Not following an applicable norm is therefore not always an act of rebellion to the authority. It is more often performed by agents who, though they recognize the authority, they believe they can justify their action according to an interpretation of the legal text, driven by legal principles, that excludes their action from the red zone of legal prohibitions. In the euthanasia case, the law prohibited to *knowingly provide the physical means or participate in the act of suicide*, with the exception of acts whose *intent is to relieve pain*. In the example, the defense argued that the law did not apply to the doctor, whose aim was to eliminate suffering, with death occurring as a mere consequence of his goal. Here, the issue concerns the interpretation of the world *pain*, and the identification of the threshold of pain that would warrant the exception. During the first trials, the interpretation of the law made by the defense seemed to match that of the jury, as the doctor was repeatedly acquitted. Interestingly, the doctor would later be found guilty of second-degree murder. In that occasion however, two things were different: first, the doctor administered the lethal injection himself; second, he publicly admitted to his deed while exposing controversial beliefs about life and religion. These circumstances highlight the link between the interpretation of the law, the goals of the action and the agent's values in the context of law application.

We note how, in both cases of *ex post* compliance checking and *ex ante* compliance choices, **the decision is twofold**: we decide our preference among *alternative representations* of the law (mostly due to *alternative interpretations*) and at the same time we decide on *a (judgment of) action*. According to Hawkins [25], there are different decisional frames to a decision. While the choice of action may depend on any of these, in *ex post* justification the legal frame prevails: the agent is expected to try to justify their action before the law, by adopting an interpretation of the law that renders their action lawful (or as little illegal as possible), even if the agent does not share that interpretation. Judges on the other hand are not supposed to be influenced by their personal preferences when adopting an interpretation: they are supposed to be impartial, not choosing according to a desired outcome, such as seeing in jail someone they despise. Still, they will apply the legal interpretation that matches the values or principles of the law as understood by them. According to Atkinson and Bench-Capon [4], the reasoning behind judgments implies an inverted *direction of fit*[11] where the judge selects the inter-

[10] See for example the Directive 2009/65/EC of the European Parliament and of the Council of 13 July 2009 on the coordination of laws, regulations and administrative provisions relating to undertakings for collective investment in transferable securities (UCITS).

[11] Discussions on principles and values have a different *direction of fit* than fact-based discussions (mind-to-world instead of world-to-mind). Notably, this allows for alternative solutions which are all equally rational [4]. The notion of *direction of fit* was introduced by John Searle in his

pretation of the norm to adapt it to its (alleged) principles. This seems to be even more the case for *ex ante* compliance choices with an *actualist* approach. In those situations, human agents do not always *conform*, i.e., adapt their actions to the law: often, they rather **adapt the interpretation of the law to the desired action**.

We have seen how, while in judgments the interpretation is aligned to the principles of the law as understood by the judge, in compliance choices the interpretation is often aligned to the subject's choice of action and justified according to values. In the next section, we outline the factors playing a role in the weighing of alternative actions in the context of a compliance choice.

4 The Factors of *ex ante* Compliance Choices

As we have seen, *ex ante* legal reasoning is shaped, to put it in the words of Leith [30], as "a process of guesswork and deductive justification". This implies that choices among interpretations are made after, and based on, the chosen action. In this latter choice, other factors come into play. Muthuri et al. [38] show how, for companies, legal choices are handled at a strategic level as a legal risk, and the interpretation of the law determines the preference among arguments in a legal dispute. Bench-Capon and Modgil [7] show how the **values** promoted/demoted by the norm play a fundamental role in determining our preference among actions as well as among alternative interpretations of a norm. For example, we do not accept an interpretation that would contravene the very principles behind the legal act that we are interpreting.

The approach of value-based reasoning seems an ideal base for any legal intelligent application dealing with *ex ante* decisions. However, as we have seen in the previous section, alternative actions and alternative interpretations are related but are not the same thing. In this section, we focus on the choice of action, trying to identify the factors that must be taken into account in order to provide a comprehensive model of compliance choices.

We identified two distinct dimensions influencing a compliance choice of action:

- The **desired state of affairs** (determined by the agent's **goals**, which in turn derive from their **value ranking** [7][12]). In the euthanasia case, the choice to assist in euthanasia would achieve a state where a person who is suffering is no more, which would fulfill the goal of reducing suffering, based on the value of caring for other individuals and freedom of individuals [44]. On the other side, a refusal would result in one more person being alive, compatible with the goals of maximising the working population; it would also avoid the creation of a caste of god-like

work on practical reasoning. He wrote: "Assume universally valid and accepted standards of rationality, assume perfectly rational agents operating with perfect information, and you will find that rational disagreement will still occur; because, for example, the rational agents are likely to have different and inconsistent values and interests, each of which may be rationally acceptable" [45].

[12] Atkinson and Bench-Capon [3] define the argument scheme for practical reasoning as "In the circumstances *R* we should perform action *A* to achieve new circumstances *S* which will realise some goal *G* which will promote some value *V*".

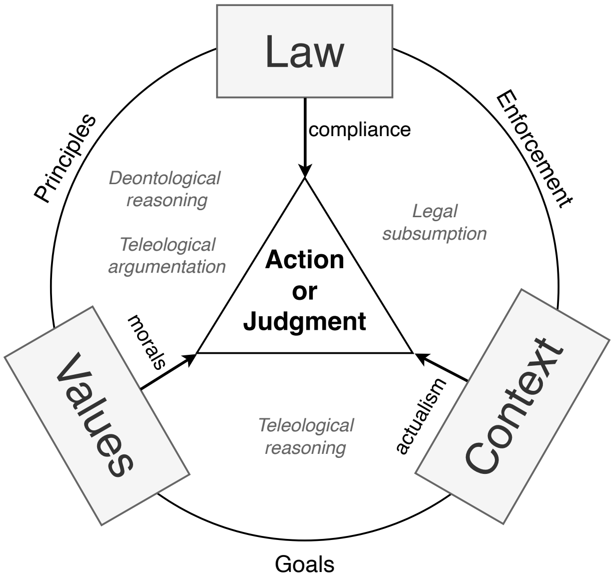

Fig. 1. The Deliberational Screwhead

purveyors of death; both of these goals pertain to the value of social security. In the AML case, the crypto firm's behaviour is aimed to have more clients, fulfilling the goal of maximizing profit, based on the value of power through resources; on the other hand, refusing transactions from sanctioned groups would result in those groups being weakened, compatible with the goal of reducing socially dangerous activities, a goal which — again — fosters the value of social security.

- The **risk** embedded in a certain behavior (mainly that of **enforcement and penalty** [41]). This factor is related to strategical thinking; however, other characteristics come into play: short of norms whose violators incur capital punishment (for which *Pascal's Wager* applies), the way in which humans see crime and punishment is a complex matter, and cannot be reduced to mere strategical thinking. As mentioned earlier, this aspect of choices is formalized in a business perspective in the discipline of risk management, and more specifically compliance risk. Regarding our examples, we can imagine how the risk of enforcement and penalty played little to no role into the doctor's choice in the euthanasia case, as he was actually looking for a prosecution in order to prove his point; at the same time, that risk/reward analysis must have played an important role in the crypto firm's choice, which later turned out to be an unlucky bet, but at the time of the decision must have seemed more likely to provide a positive economic outcome than a negative one.

The diagram in Fig. 1 represents the factors interacting in a compliance choice, which we call the *Deliberational Screwhead*; its main elements, as identified by Verheij [48] in the context of ethical decision making, are *law*, *context* and *values*.

- ***Law*** is intended as the sum of statutes, case-law, guidelines, and any other official document expressing norms. Resolution of legal conflicts and other formal legal reasoning is performed here, before the result in terms of applicable norms is carried on to influence the choice of action or judgement. Law influences the choice of *judgment* or *action* through ***compliance***: to the extent to which an action (or judgement) is determined by the law, we can say that it is influenced by compliance.
- ***Context*** is intended as the contingent state of affairs (the facts) to be seen either as *case* or *situation* [17], depending on whether we are discussing *ex post* or *ex ante* (i.e., whether we decide on a *judgment* or an *action*). Context influences the choice of *judgment* or *action* through ***actualism***: to the extent to which an action or judgement is determined by the context, we can say that it is influenced by actualism.
- ***Values*** are, in the words of Schwartz [44], "conceptions on the desirable that influence the way people select action and evaluate events". They influence a compliance decision through the so-called value ranking [4]. As noted previously, in the euthanasia case, the ranking would prioritize freedom of individuals in the case of a pro-euthanasia behavior, and the preservation of human life in the case of a contrary position. In Schwartz's value theory [44], a pro-euthanasia approach seems to prioritize *self direction (action)* over *societal security*, while the opposite applies to the contrary position. In the AML example, the values involved seem to be *power (resources)* vs. *societal security*, *norm conformance*, and *face* (i.e., avoiding the loss of reputation in case the breach was discovered). Unfortunately, we are currently unable to specify human goals and societal values in a way that reliably directs AI behavior [39]. Values influence the choice of *judgment* or *action* through ***morals***: to the extent to which an action or judgement is determined by their values, we can say that it is influenced by morals.

A choice is hardly ever influenced by a single factor. In that sense, we can see how some concepts of legal compliance such as ***legal principles***, ***goals*** and ***enforcement*** can be seen as a combination of factors (law and values, values and context, and context and law, respectively). Furthermore, these combinations of factors seem to imply different types of reasoning: reasoning on principles is related to deontological reasoning and teleological argumentation; reasoning on goals is related to teleological reasoning; and reasoning on enforcement is related to legal subsumption, a type of legal interpretation. An important question here is whether (and to what extent) these factors of compliance choices should be made explicit in the deliberation process. This question is related to whether we consider legal compliance as being limited to detect *rule conformance* or as encompassing *rule following*[13].

The factors that we have outlined in this section work as prioritization criteria among alternatives of action. In the next section, we will investigate how such prioritization criteria can be implemented in an argumentation framework through a meta-argumentation layer.

[13] While in *rule conformance* we limit our analysis to the compliance of the behavior with the norm, in *rule following* we also want to verify the congruence of the agent's intentions with the legal principles. The difference is important when drawing insights from observing an agent's behavior — see Galoob and Hill [21]. The first to argue for rule following was Kant [28].

5 Handling Compliance Decisions with Argumentation

The goal of Dung's abstract argumentation framework is to determine the acceptability of arguments. The framework is composed of only two elements: arguments and binary conflict-based attack relations [18]. For example, in the euthanasia case, the argument that "*the doctor acted with the purpose of economic profit*" is attacked by the argument that "*the doctor received no economic compensation for his action*". Dung's abstract argumentation framework does not include a solution to solve conflicts between argument, i.e., arguments that attack each other reciprocally. For example the arguments "*causing death to remove pain is not murder*" and "*causing death is always murder*" are in conflict. To handle such conflicts, an **extended argumentation framework** with a meta-argumentation layer is necessary [12]. Modgil [35] proposed such a framework for value-based reasoning, prioritizing alternative interpretations of a norm according to the value ranking of the agent. Despite the framework being "usable in any situation in which reasoning about actions is required" (Atkinson and Bench-Capon [4]), when dealing with law as a domain-specific application, the literature focuses on *ex post* judgments, showing how value rankings lead to the preference among alternative interpretations of the law[14].

In the context of our example, Fig. 2 shows the arguments involved in the euthanasia case. The white circles represent arguments, and the arrows represent attack relations. Compared to Dung's abstract framework [18], we added a grey circle representing a meta-argument via value ranking. On the left-hand side, we see the conflict between the arguments "*causing death to remove pain is not murder*" and "*causing death is always murder*". Here, the conflict can be resolved via a value ranking between social security and personal freedom. On the right-hand side, the value-based conflict is shown in details, following Modgil's extended argumentation framework [36]. Here, the two white circles represent the conflicting arguments. Their attack relations are themselves attacked by the two possible value preference arguments (grey circles), namely *v2>v1* and *v1>v2*. Such value-based conflict is solved in a specific extension of the framework by determining the ranking of *v1* and *v2*.

The application of practical reasoning to moral choices is investigated by Atkinson and Bench-Capon [3], but without taking into account the enforcement risk, and has not been adopted beyond theoretical research. Nevertheless, the framework for value-based reasoning seems fit for representing the inner deliberation of *ex ante* compliance choices: different value rankings would lead to different (and equally rational) choices of action and preferences among alternative interpretations. For example, we may see how choosing a prohibited action could derive from a deliberation involving a preference of value *v2* over value *v1*, which made the agent prefer an action promoting value *v2* while the norm seems to promote value *v1*; we could then verify whether it exists an interpretation of the breached norm under the same value ranking *v2>v1* which would result in the same action being compliant to the norm.

The value-based conflicts in the euthanasia case are provided in Fig. 3. We can assume the desire of avoiding conviction as the main argument attacking the choice of

[14] In Atkinson and Bench-Capon [4], the example concerns the analogy in case law, a typical use case for AI and Law in common law systems.

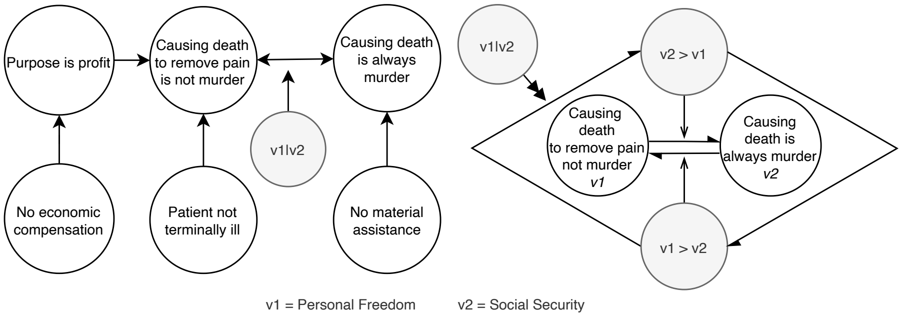
Purpose is profit
Causing death to remove pain is not murder
Causing death is always murder
v1Iv2
No economic compensation
Patient not terminally ill
No material assistance
v1Iv2
v2 > v1
Causing death to remove pain not murder v1
Causing death is always murder v2
v1 > v2
v1 = Personal Freedom
v2 = Social Security

Fig. 2. *Left*: possible representation of the *ex post* judgment on the euthanasia case using Dung's abstract framework with the addition of meta-arguments (grey circle); *right*: value-based conflict in an extended argumentation framework, as given by Modgil and Bench-Capon [36].

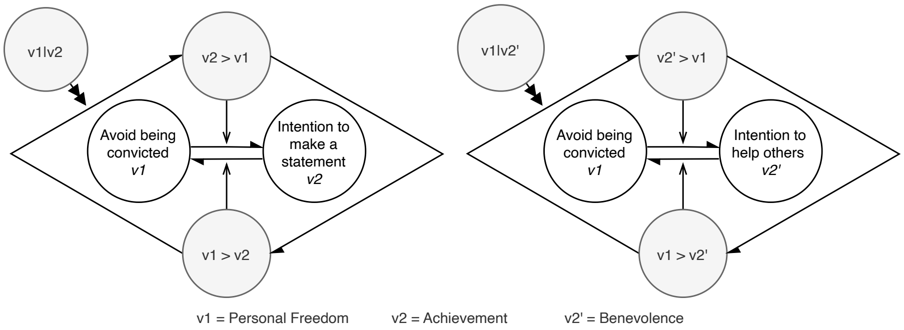

Fig. 3. Possible value-based conflicts in the *ex ante* choice on euthanasia.

assisting in euthanasia. This argument is in conflict with two possible intentions of the doctor, i.e., that of helping others and that of making a public statement and possibly gain renown. These conflicts can be solved via meta-arguments referring to the value ranking and the enforcement risk; however, in the example, the doctor was undeterred by the latter: he seemed convinced of his capability to avoid punishment, or ready to accept it, thus both instances of *v2* are preferred over *v1* in this case[15]. This is what characterizes, when referring to the diagram in Fig. 1, the choice of the doctor as a *moral* (rather than a *compliant* or *actualist*) choice. The two diagrams in Figs. 2 and 3 are related in that, after a choice in the inner deliberation of Fig. 3 has been taken, the argumentation of Fig. 2 represents the way in which the doctor (more precisely his attorney) will argue for the lawfulness of his (client's) action. Only a combination of

[15] The value of *personal freedom* leads to an *ex post* judgment of euthanasia as lawful, while in an *ex ante* choice leads to abstaining from such an act. This difference further explains the importance of accurately modeling the factors of compliance decisions in both scenarios.

the two diagrams can in fact explain the inter-relationship between choice of action and preference among interpretations (see § 3.3).

It seems that argumentation has a different task in the case of compliance choices: rather than to verify a complete set of facts (a *case* [17]) under alternative interpretations, the task concerns the analysis of different choices of action combined with different interpretations of the law in a specific *situation*. On that regard, ambiguity propagation logic [24] allows for exploring alternative conclusions in the presence of unknown or uncertain information on the acceptability of specific arguments.

6 Conclusions

In this paper, we have investigated the representation of compliance decisions from the perspective of both *digital judges* and *digital agents*. We discussed the state of the art in the representation of *ex ante* and *ex post* compliance decisions and tried to expand on the factors that lead to a *compliance choice*. We saw how compliance choices are related to different factors and how they influence the preference among alternative interpretations of a norm. While value-based argumentation [4] seems able to support *ex ante* reasoning, this framework is not adopted in practical applications, where instead a naive approach to compliance choices prevails, with agents supposed to spontaneously comply with all the rules applicable to their situation.

A fit model for digital judges would perform poorly for a digital agent if its motivations of action were built solely upon the justifications provided in court, ignoring the inner deliberation of *ex ante* compliance choices, and the relation between the choice of action and the (declared) preference among alternative interpretations. We showed how, in our euthanasia example, the two argumentations are inherently different.

From this consideration arises an important concern on AI alignment [16]: automated agents should make compliance choices by contextualizing value(-based goal)s and normative restrictions in the specific situation, as a human would. In this paper, we have outlined how context and values shape not only the legal interpretation of the rules, but also the uncertainties of ethical decision making [48], which affect the utility functions of practical reasoning [14]. We argue that the research towards a general legal architecture for intelligent applications should extend legal case-based reasoning [3] and value-based argumentation [4] to encompass practical reasoning and game-theoretical aspects of *ex ante*, *actualist* compliance choices [11].

We need to represent human legal reasoning (i.e., practical reasoning) as it is, and not as we would like it to be (i.e., mere rules application), to avoid the risk that biased AIs, unaware of our inner deliberation due to an inaccurate representation of legal reasoning, reproduce an *ex post* bias which, in a feedback loop [40], has the potential to downplay the relationship of humans with law towards thoughtless rule conformance.

Acknowledgments

This research was funded in whole, or in part, by the Luxembourg National Research Fund (FNR), under grant number C24/IS/18894115/AGLAIA.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of data and analytics, pp. 254–264. Auerbach Publications (2022)
2. Anim, J., Robaldo, L., Wyner, A.Z.: A SHACL-based approach for enhancing automated compliance checking with RDF data. Information **15**(12) (2024)
3. Atkinson, K., Bench-Capon, T.: Addressing moral problems through practical reasoning. J. Applied Logic **6**, 135–151 (01 2008)
4. Atkinson, K., Bench-Capon, T.J.: Value-based argumentation. FLAP **8**(6), 1543–1588 (2021)
5. Bahl, U., Topaz, C., Obermuller, L., Goldstein, S., Sneirson, M.: Algorithms in judges' hands: Incarceration and inequity in Broward County, Florida. UCLA L. Rev. Disc. **71** (2023)
6. Bench-Capon, T.: Before and after Dung: Argumentation in AI and Law. Argument & Computation **11**(1-2), 221–238 (2020)
7. Bench-Capon, T., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artif. Intell. Law **25**(1), 29–64 (Mar 2017)
8. van Berkel, K., Straßer, C.: Reasoning with and about norms in logical argumentation. In: Computational Models of Argument. vol. 353. IOS Press (2022)
9. Berman, D.H., Hafner, C.D.: Representing teleological structure in case-based legal reasoning: the missing link. In: Proceedings of the 4th Int. Conf. in AI and Law. ICAIL '93, Association for Computing Machinery, New York, NY, USA (1993)
10. Billi, M., Calegari, R., Contissa, G., Lagioia, F., Pisano, G., Sartor, G., Sartor, G.: Argumentation and defeasible reasoning in the law. J — Multidisciplinary Scientific Journal **4**, 897–914 (2021)
11. Boella, G., van Der Torre, L.: A game-theoretic approach to normative multi-agent systems. Normative Multi-agent Systems (2007)
12. Boella, G., Gabbay, D.M., van der Torre, L., Villata, S.: Meta-argumentation modelling I: Methodology and techniques. Studia Logica **93**(2), 297 (2009)
13. Boella, G., Humphreys, L., Muthuri, R., Rossi, P., van der Torre, L.: A critical analysis of legal requirements engineering from the perspective of legal practice. In: RELAW 2014 (2014)
14. Bradley, R., Drechsler, M.: Types of uncertainty. Erkenntnis **79**(6), 1225–1248 (2014)
15. Breaux, T.D., Antón, A.I., Doyle, J.: Semantic parameterization: A process for modeling domain descriptions. ACM Trans. Softw. Eng. Methodol. **18**(2), 5:1–5:27 (2008)
16. Christian, B.: The alignment problem: How can machines learn human values? Atlantic Books (2021)
17. Čyras, V., Lachmayer, F.: Situation versus case and two kinds of legal subsumption. In: Abstraction and Application, Proc. of the 16th Int. Legal Informatics Symposium, IRIS (2013)
18. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2) (1995)
19. Esayas, S., Mahler, T.: Modelling compliance risk: a structured approach. Artificial Intelligence and Law **23**(3), 271–300 (2015)
20. Francesconi, E., Governatori, G.: Patterns for legal compliance checking in a decidable framework of linked open data. Artificial Intelligence and Law **31**, 1–20 (2022)
21. Galoob, S.R., Hill, A.: Norms, attitudes and compliance. Tulsa Law Review **50**(2) (2015)
22. Gervais, D.J., Nay, J.J.: Artificial intelligence and interspecific law. Science **382**(6669), 376–378 (2023)
23. Ghanavati, S., Amyot, D., Rifaut, A.: Legal goal-oriented requirement language (legal GRL) for modeling regulations. In: MiSE 2014. pp. 1–6 (2014)
24. Governatori, G., Maher, M.J., Antoniou, G., Billington, D.: Argumentation semantics for defeasible logic. Journal of Logic and Computation **14**(5), 675–702 (2004)

25. Hawkins, K.: Enforcing regulation: Working theories of compliance and punishment. In: Proc. of the Conference on the Enforcement of Regulation; Indecopi: Lima, Peru (2016)
26. Hellmeier, S., Cole, R., Grahn, S., Kolvani, P., Lachapelle, J., Lührmann, A., Maerz, S.F., Pillai, S., Lindberg, S.I.: State of the world 2020: autocratization turns viral. Democratization **28**(6), 1053–1074 (2021)
27. Ingolfo, S., Siena, A., Susi, A., Perini, A., Mylopoulos, J.: Modeling laws with nomos 2. In: RELAW 2013. pp. 69–71 (2013)
28. Kant, I.: The Metaphysics of Morals. Cambridge University Press (1797)
29. Kosenkov, O., Unterkalmsteiner, M., Méndez, D., Fucci, D., Gorschek, T., Fischbach, J.: On developing an artifact-based approach to regulatory requirements engineering. In: MoDRE 2024. pp. 262–271 (2024)
30. Leith, P.: Logic, formal models and legal reasoning. Jurimetrics J. **24**, 334 (1983)
31. Longo, L.: Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. ML for Health Informatics (2016)
32. Luger, E., Moran, S., Rodden, T.: Consent for all: revealing the hidden complexity of terms and conditions. In: Proc. SIGCHI Conference on Human Factors in Computing Systems. p. 2687–2696. CHI '13, Association for Computing Machinery, New York, NY, USA (2013)
33. Makinson, D., Van Der Torre, L.: Constraints for input/output logics. Journal of philosophical logic **30**, 155–185 (2001)
34. de Mattos, L.A., Rocha, R., de Castro, F.: Human error and violation of rules in industrial safety: A systematic literature review. WORK **79**(3), 1237–1253 (2024)
35. Modgil, S.: Reasoning about preferences in argumentation frameworks. Artificial Intelligence **173**(9-10), 901–934 (2009)
36. Modgil, S., Bench-Capon, T.: Integrating object and meta-level value based argumentation. COMMA **172**, 240–251 (2008)
37. Modgil, S., Prakken, H.: A general account of argumentation with preferences. Artificial Intelligence **195**, 361–397 (2013)
38. Muthuri, R., Capecchi, S., Sulis, E., Amantea, I.A., Boella, G.: Integrating value modeling and legal risk management: an IT case study. Information Systems and e-Business Management pp. 1–29 (2022)
39. Nay, J.J.: Law informs code: A legal informatics approach to aligning artificial intelligence with humans. Nw. J. Tech. & Intell. Prop. **20**, 309 (2022)
40. O'Hara, I.: Feedback loops: Algorithmic authority, emergent biases, and implications for information literacy. Pennsylvania Libraries: Research & Practice **9**(1), 8–15 (2021)
41. Peeters, M., Denkers, A., Huisman, W.: Rule violations by SMEs: The influence of conduct within the industry, company culture and personal motives. European Journal of Criminology **17**(1), 50–69 (2020)
42. Raz, J. (ed.): Practical Reasoning. Oxford University Press, New York (1978)
43. Rotolo, A.: Norm compliance of rule-based cognitive agents. In: Proceedings of the 22nd International Joint Conference on AI - Volume 3. IJCAI'11, AAAI Press (2011)
44. Schwartz, S.H.: The refined theory of basic values. In: Values and behavior: Taking a cross cultural perspective, pp. 51–72. Springer (2017)
45. Searle, J.R.: Rationality in action. MIT press (2003)
46. Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L.C., Ceci, M., Dann, J.: An automated framework for the extraction of semantic legal metadata from legal texts. Empir. Softw. Eng. **26**(3), 43 (2021)
47. Steinfeld, N.: "I agree to the terms and conditions": (How) do users read privacy policies online? An eye-tracking experiment. Computers in Human Behavior **55**, 992–1000 (2016)
48. Verheij, B.: Formalizing value-guided argumentation for ethical systems design. Artif. Intell. Law **24**(4), 387–407 (Dec 2016)
49. Walton, D., Reed, C., Macagno, F.: Argumentation schemes. Cambridge Univ. Press (2008)