Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-171

A Computational Model of Moral and Legal Responsibility via Simplicity Theory

Giovanni SILENO [a,1], Antoine SAILLENFEST [b] and Jean-Louis DESSALLES [a]
[a] *LTCI, Télécom ParisTech, Université Paris-Saclay, 46 rue Barrault, Paris, France*
[b] *Geronimo Agency, 33 rue d'Artois, Paris, France*

Abstract. Responsibility, as referred to in everyday life, as explored in moral philosophy and debated in jurisprudence, is a multiform, ill-defined but inescapable notion for reasoning about actions. Its presence in all social constructs suggests the existence of an underlying cognitive base. Following this hypothesis, and building upon *simplicity theory*, the paper proposes a novel computational approach.

Keywords. moral responsibility, legal responsibility, simplicity theory, foreeseability, inadvertence, risk, negligence

1. Introduction

The notion of *individual responsibility* is paramount in informal social relationships as much as in formal legal institutions. With the (supposedly) near advent of autonomous entities, its formalization becomes a pressing problem. In human societies, responsibility attribution is a spontaneous and seemingly universal behaviour. Non related ancient legal systems (e.g. [9]) bear much resemblance to modern law and seem perfectly sensible nowadays. This universality suggests that responsibility attribution may be controlled, at least in part, by fundamental cognitive mechanisms. Experimental studies showed that various parameters influence moral responsibility attribution [15]. For instance, people are more prone to blame (praise) an agent for an action if they are closer to the victims (beneficiaries), if the outcome follows in a simple way from the action or if the agent was able to foresee the outcome. Several of these parameters, such as the agent's foreseeing ability, are purely cognitive. Theories of law indeed take cognition into account with notions such as *mens rea*. Following this idea, the present paper attempts to bridge the gap between cognitive modelling and theory of law.

The AI & Law literature proposes two main approaches to *responsibility attribution*. The structural approach attempts to capture reasoning constructs using ontologies [10], inference [14] or stories [1]. The probabilistic approach focuses on quantifying the relative support of evidence in the reasoning process, e.g. via Bayesian inference [6] or causal Bayesian networks [7,3]. Hybrid proposals exist as well [17]. The present work introduces an alternative framework, using notions from *simplicity theory* [4], offering a potential ground for unification: because *simplicity theory* relies on the computation of Kolmogorov-like complexities, it involves both structural and quantitative aspects.

The paper proceeds as follows. In § 2, we consider a few accounts of the notion of responsibility with some case examples. In § 3, we briefly introduce *simplicity theory* and show how it can deal with moral evaluation. In § 4, we evaluate it based on the given case examples. A note on further developments ends the paper.

[1] Corresponding author: giovanni.sileno@telecom-paristech.fr

2. Causal, Legal and Moral Responsibilities

For some legal scholars, a theory of responsibility should rely on legal causation rather than on factual causation (see the overviews on causal minimalism given in [10], [12]). Indeed, the Greek word for "cause" started as a legal term [11]. Consider this real case:

Example 1. (Two bad hunters) *Two hunters negligently fired their shotguns in the direction of their guide, and a pellet lodged in his eye. Because it was impossible to tell which hunter fired the shot that caused the injury, the court held both hunters liable.*[2]

Here, one of the two hunters is held responsible despite the fact that he did not materially cause the damage. Physical causation is rarely matter of dispute, and when it is, as in the previous case, it is irrelevant to formulate a legal judgement. By contrast, legal causation is always relevant and is much debated when attributing responsibility (with variations depending on the different legal traditions). Consider the following case:

Example 2. (Navigating oil) *At a landing stage, furnace oil spilled into the water for defendants' negligence. The oil spread on the water surface, reaching a nearby ship on which welding work was being carried out. Sparks ignited the oil, which caught on fire damaging several vessels. The court held that contamination damage caused by the oil was reasonably foreseeable, but that damage caused by fire was not foreseeable and was thus too remote for recovery.*[3]

The core of the dispute was to settle on *foreseeability*, i.e. the ability to predict the consequences of an event or action. Beyond foreseeability, events would be too *remote* to the defendant to be accounted liable for, even if they were enabling the actual chain of causation. Although foreseeability is a fictional device, knowing what-caused-what or what-enabled-what—*pace* causal minimalists—influences its evaluation:

Example 3. (Navigating oil, cont'd) *Further evidence revealed the presence of floating flammable objects in the water which, combined with the oil, made the lightning of the fire more probable. The court held the defendant liable, because, seen the magnitude of the risk, a reasonable person would have reacted to prevent it.*[4]

The second judgement not only considers the ability to foresee alternative causal chains, but also takes the magnitude of the risk into account. However, not every responsibility attribution is about the agents' rational abilities. Consider this simple case:

Example 4. (A broken vase) *A person enters in a shop and breaks inadvertently a vase. According to the law, she is usually liable to provide compensation, but not to be blamed.*

Even when people are making reasonable choices, things may go wrong. These cases are usually under the scope of law (but not necessarily of morality), in order to apply a fairer redistribution of the losses amongst the parties (principle of *equity*).

Legal Responsibility and its Boundaries Legal systems usually have distinct mechanisms to decide on liability (*who has to provide remedy?*, as in the previous examples) and on blame (*who has to be punished?*). In general, guiltiness is attributed by proving a combination of factual elements under the scope of law (*actus reus*) and mental elements relevant to the case (*mens rea*). Consider however this famous paradox [13, Ch. 10]:

[2] Summers v. Tice (1948), 33 Cal.2d 80, 199 P.2d 1.
[3] Overseas Tankship (UK) Ltd v. Morts Dock and Eng. Co Ltd or "*Wagon Mound (No. 1)*" (1961), UKPC 2.
[4] Overseas Tankship (UK) Ltd v The Miller Steamship Co or "*Wagon Mound (No. 2)*" (1967), 1 AC 617.

Example 5. (The desert traveller). *A desert traveller T has two enemies. Enemy 1 poisons T's canteen and Enemy 2, unaware of Enemy 1's action, empties the canteen. A week later, T is found dead and the two enemies confess to action and intention. It is then discovered that T never drank from the canteen and died by dehydration.*

From a causal point of view, this example contains a *pre-emption*: an event prevents another event from being successful. Is Enemy 1 guilty? In principle, law disregards potential outcomes, so the answer is no. Intuitively, however, Enemy 1 is morally guilty. And many legal systems do attribute some charge to the offender who willingly initiated a course of action that may have lead to a crime (e.g. *attempted murder*).

3. Theoretical Framework

This section briefly presents *simplicity theory* (ST) as a theoretical basis to construct computational models of judgement. ST is a cognitive theory stemming from the observation that human individuals are highly sensitive to *complexity drops* [4]: i.e. to situations that are *simpler to describe than to explain*. The theory builds on notions and tools from *algorithmic information theory* (AIT) that are redefined with respect to cognitive agents. It has been used to make predictions, confirmed empirically, about what humans would regard as *unexpected*, *improbable*, and *interesting* [5,15,16].[5]

Unexpectedness A central notion in ST is *unexpectedness* (U), defined as:

$$U(s) = C_W(s) - C_D(s) \tag{1}$$

where s is a situation, $C_W(s)$ is the complexity of the circumstances that were necessary to generate s, $C_D(s)$ is the complexity of describing s. The two complexities are versions of *Kolmogorov's complexity*, which, informally, is the length in bits of the shortest description of an object. ST distinguishes *causal* complexity ($C_W(s)$) from the usual *description* complexity ($C_D(s)$). Determining a causal path requires adding the complexities of making a choice at successive choice points. If there are k equivalent options at a choice point, one needs $\log_2(k)$ bits to make a decision. On many occasions, $C_W(s)$ corresponds to the logarithm of the probability of occurrence. Complexity computations, however, have a broader range of applicability, as for instance when dealing with unique events. Using $C_W(s)$ we can define the *causal contribution* of a situation s_1 to bringing about a second situation s_2:

$$R(s_1, s_2) = C_W(s_2) - C_W(s_2||s_1) \tag{2}$$

where $C_W(s_2||s_1)$ is the complexity of causally generating s_2, starting from a state of the world in which s_1 holds. If $R(s_1, s_2) = 0$, the two events are independent. If $R(s_1, s_2) > 0$ (respectively < 0), s_1 concurs positively (negatively) to the occurrence of s_2.

The description complexity $C_D(s)$ specifies the shortest *determination* of an object s. For instance, the shortest determination of s may consist in merely retrieving it from memory (think of referring to famous people). In this case, $C_D(s)$ amounts to the complexity of the parameter controlling the retrieval, i.e., considering memory as an ordered set, the $\log_2$ of the index of the object in that set (frequently used objects have smaller indexes). Applying similar considerations to spatio-temporal properties, we observe that C_D captures the distance (as inverse of proximity) of the agent to the situation.

[5] For a general presentation see: `http://simplicitytheory.org`.

Points of View For any agent A, C_W^A will denote the generation complexity computed by A using her knowledge. Different *points of view* may lead to alternative computations of causal complexity for the same situation.

Emotion and Intention Unexpectedness captures the epistemic side of a *relevant* experience. For the *epithymic* (i.e. concerning desires) side, ST refers to a representation of *emotion* limited to considering intensity E and valence ε. Focusing only on intensity, we define the *actualized* (or *hypothetical*) *emotion* as $E_h(s) = E(s) - U(s)$, pruning the emotion of its unexpectedness[6]. Intention is driven by E_h^A, computed from the point of view of an agent A who considers performing action a. If A sees a as the shortest causal path to s, $U^A(s) = U^A(a) + U^A(s||a)$, and intention turns out to be:

$$I(a) = E^A(s) - U^A(s||a) - U^A(a) \tag{3}$$

When a is intended (volitional), $U(a) = 0$. This term, when non-zero, represents ***inadvertence***. Note that in the more general case, intention should result from an aggregation of similar components for different outcomes s_i.

Moral Responsibility and Judgement Our central claim is that the difference between intention and of moral responsibility is one of *point of views*. To obtain intention, we consider the point of view of the actor A for all the components. When performing moral evaluation, however, the observer applies her own point of view (we omit superscript O), except for the elements concerning the action, which are computed using her *model* of the actor. The ***moral responsibility*** M attributed to A by observer O is defined as:

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a) \tag{4}$$

The superscript $^{\downarrow A}$ means that O uses her model of A to compute U (e.g. a prescribed role, a reasonable standard, etc.). If we introduce the actualized emotion term we have: $M(a) = E_h(s) + U(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$, from which, making C_W and C_D explicit, we can extract the ***causal responsibility*** component:

$$R^{\downarrow A}(a,s) = C_W(s) - C_W^{\downarrow A}(s||a) \tag{5}$$

This formula captures how much A's action a was supposed to bring about s in A's mind. If we suppose than $C_D^{\downarrow A}(s|a) \approx 0$ — a simplification possible when the conceptual relation between cause and effect is proximate (i.e. in A's model, the action is directly linked to the outcome) — the resulting equation is:

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a) \tag{6}$$

In words, the intensity of moral evaluation increases with the *actualized emotional intensity* and with *causal responsibility*, decreases with the *remoteness* of the consequence to the observer (proximate situations are simpler to describe) and with *inadvertence*.[7]

Now, imagine the case of a famous singer who is killed as a casual bystander in a car accident. The popular emotion might be so strong that the police have to save the car

[6] In a utilitarian perspective, E_h may be interpreted as the logarithmic version of the expected value, and E as the logarithm of the absolute value of gain or loss.

[7] Like for intention, a complete *moral judgement* of a positive action a should take into account also the evaluation of its *omission*, in order to capture e.g. the fact that someone may act negatively to avoid even worst consequences (cf. attenuating circumstances).

driver from being lynched. An impartial judge must consider the victim as if she were any person. This means that *equality* in judgement is obtained by reducing the impact of C_D, i.e. by *recomplexifying* the mental simplification due to proximity effects.

4. Applying Simplicity Theory to Judgement

We now examine how the framework presented above matches our examples.

Two bad hunters Two hunters (A_1, A_2) fire *negligently* at their guide (a_1, a_2), resulting in his injury (s). Causal contributions—$R(a_1,s)$ or $R(a_2,s)$—cannot be determined. Negligence is captured when actors fail to *foresee* the unlawful consequences of their action: $C_W^{A_1}(s||a_1) = C_W^{A_2}(s||a_2) \gg 0$. However, it is reasonable to expect that the two actions may have resulted in that outcome (note that $C_W(s) \gg 0$): $C_W^{\downarrow A_1}(s||a_1) = C_W^{\downarrow A_2}(s||a_2) > 0$ and $R^{\downarrow A_1}(a_1,s) = R^{\downarrow A_2}(a_2,s) > 0$. Therefore, both hunters receive the same moral evaluation. Generalizing this case, the ***negligence*** of an actor A for an action a w.r.t. a consequence s is defined as:

$$N^A(a,s) = C_W^A(s||a) - C_W^{\downarrow A}(s||a) \tag{7}$$

Navigating oil The oil leakage at the landing stage (s_1) results from an omission of adequate care $(a = \neg b)$ by defendant A. The case centers around responsibility attribution for the fire at the near wharf (s_2). The court held that though s_1 was foreseeable, s_2 was not: $C_W^{\downarrow A}(s_1||a) \sim 0$ and $C_W^{\downarrow A}(s_2||s_1) \gg 0$, and $R(s_1,s_2) \sim 0$. Integrating the C_D terms, we define A's ***foreseeability*** of the consequence s of an action a as negated unexpectedness:

$$F^A(a,s) = -U^{\downarrow A}(s||a) \tag{8}$$

(F^A is in $]-\infty, 0]$, 2^{F^A} in $[0,1]$; $F^A = 0$, $2^{F^A} = 1$ when s is perfectly foreseeable after a.)

Navigating oil, cont'd Due to the presence of flammable objects (s_1'), the defendant should have reasonably anticipated the consequences: $C_W^{\downarrow A}(s_2||a \wedge s_1) > C_W^{\downarrow A}(s_2||a \wedge s_1 \wedge s_1')$. Foreeseability increases, and so does responsibility. The court made also an argument about weighting of risks. Traditionally, risks are approached with *expected value*. Considering $E(s)$ as the "win" value (loss in this case), the ***risk*** can be defined as:

$$K^A(a,s) = E(s) - U^{\downarrow A}(s||a) = E(s) + F^A(a,s) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) \tag{9}$$

This view agrees with Hart and Honoré's [8] consideration of risk as a generalization of foreseeability, providing an *upper bound* for the damages to be paid.

A broken vase A person A slips in a shop (a) and breaks a vase (s). For a person to slip is unexpected but still possible: $U(a) > 0$ with a good probability of breaking something ($U^A(s||a) \sim 0$, $C_W(s||a) > 0$ and $R(a,s) \gg 0$). We get: $M(a) \approx E(s) - U^{\downarrow A}(a)$. This expression accounts for the fact that the agent and the shopkeeper may have different evaluations of $M(a)$, due to their different appraisal of $E(s)$.

The desert traveller Enemy 1 (E_1) poisons the canteen (a_1); Enemy 2 (E_2) empties the canteen (a_2). Instead of getting poisoned (s_1), the desert traveller gets dehydrated (s_2) and dies (s). We have: $C_W(s) \gg 0$, $C_W(s_1||a_1) = C_W(s_2||a_2) = C_W(s||s_2) = C_W(s||s_1) = 0$, and $C_W(s_2||a_1) \gg C_W(s_2||a_2) = 0$. Then, $R(a_1,s_2) = 0$, but also $C_W(s||s_2) - C_W(s||s_2 \wedge a_1) = 0$, which explains why E_1 is not judged causally responsible for the occurrence of s, knowing that s_2 was the case. However, $R^{\downarrow E1}(a_1,s) \gg 0$, which explains why E_1 is regarded as morally responsible (Eq. (6)).

5. Conclusion and Further Developments

The hypothesis advanced here is that moral and legal responsibility attributions share a fundamentally similar cognitive architecture. We could derive from *simplicity theory* formal definitions of: *intention* (3), *moral responsibility* (4, 6), *causal responsibility* (5), *inadvertence*, *negligence* (7), *foreseeability* (8), *risk* (9). These results are however preliminary, and further investigation is needed to compare them with existing proposals (see § 1). For instance, the analytic definitions of *degree of responsibility* and *blame* given in [7] are aligned with those of *causal contribution* (2) and *causal responsibility* (5).

As observed in the domain of legal ontologies [2], legal reasoning builds upon *normative knowledge* (qualifying behaviour as allowed and disallowed) and *responsibility knowledge* (assigning responsibility for the behaviour). The former is fed mostly by world definitional knowledge, the second by world causal knowledge. Our model is aligned with this analysis, for the crucial role of world complexity (C_W). For its cognitive flavour, our proposal offers an alternative contribution on responsibility in the field of AI and Law. Furthermore, for its grounding on Kolmogorov complexity, it offers a computational alternative to probability-based approaches (e.g. [6]), not requiring the reference to *a priori* probabilities, but referring to cognitively grounded elements. The richness of the framework opens new spaces for further interaction with legal analysis, analytic proposals, and for comparisons with empirical results.

References

[1] F. J. Bex, P. J. Van Koppen, H. Prakken, and B. Verheij. A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law*, 18(2):123–152, 2010.
[2] J. A. Breuker and R. G. F. Winkels. Use and Reuse of Ontologies in Legal Knowledge Engineering and Information Management. *Proc. of Int. Workshop on Legal Ontologies (LegOnt'03)*, 2003.
[3] H. Chockler, N. Fenton, J. Keppens, and D. A. Lagnado. Causal analysis for attributing responsibility in legal cases. *Proc. of 15th Int. Conf. on Artificial Intelligence and Law (ICAIL15)*, pages 33–42, 2015.
[4] J. L. Dessalles. Algorithmic simplicity and relevance. *Algorithmic probability and friends*, 7070 LNAI:119–130, 2013.
[5] A. Dimulescu and J.-L. Dessalles. Understanding Narrative Interest : Some Evidence on the Role of Unexpectedness. *Proc. of 31st Conf. of the Cognitive Science Society*, pages 1734–1739, 2009.
[6] N. Fenton, M. Neil, and D. a. Lagnado. A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, 37(1):61–102, 2012.
[7] J. Y. Halpern. Cause, responsibility and blame: A structural-model approach. *Law, Probability and Risk*, 14(2):91–118, 2015.
[8] H. L. A. Hart and T. Honoré. *Causation in the Law*. Clarendon Press, 1985.
[9] U. Lau and T. Staack. *Legal Practice in the Formative Stages of the Chinese Empire*. Brill, 2016.
[10] J. Lehmann, J. A. Breuker, and P. W. Brouwer. Causation in AI & Law. *Artificial Intelligence and Law*, 12(4):279–315, 2004.
[11] R. McKeon. The Development and The Significance of the Concept of Responsibility. *Revue Int.e De Philosophie*, 11(39):3–32, 1957.
[12] M. S. Moore. *Causation and Responsibility*. Oxford University Press, 2009.
[13] J. Pearl. *Causality*. Cambridge University Press, 2009.
[14] H. Prakken. An exercise in formalising teleological case-based reasoning. *Artificial Intelligence and Law*, pages 49–57, 2002.
[15] A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. *Proc. of 34th Conf. of the Cognitive Science Society*, pages 947–952, 2012.
[16] A. Saillenfest and J.-L. Dessalles. Some Probability Judgments may Rely on Complexity Assessments. *Proc. of 37th Conf. of the Cognitive Science Society*, pages 2069–2074, 2015.
[17] B. Verheij. To catch a thief with and without numbers: Arguments, scenarios and probabilities in evidential reasoning. *Law, Probability and Risk*, 13(3-4):307–325, 2014.