Some challenges for formal theories of responsibility

Ilgar Gapagov* and Matteo Pascucci*

* Department of Philosophy, CEU, Vienna, Austria
Gapagov_Ilgar@student.ceu.edu, PascucciM@ceu.edu

Abstract

This article is dedicated to the analysis of the notion of responsibility in logic. Two fundamental conditions are involved in an ascription of moral or legal responsibility: the control condition and the awareness condition. We focus on some neglected aspects of these conditions by presenting three scenarios that pose challenges to existing accounts and by offering a formalization of these scenarios in a new framework of multimodal logic.

1 Introduction

Many logic accounts of responsibility are inspired by Fischer and Ravizza's theory [4], which revolves around two main conditions: that the agents to whom we want to ascribe responsibility had sufficient control over the outcome (*control condition*), and that they were sufficiently aware of the connections between their behaviour and the outcome (*awareness condition*). These two conditions can be traced back to Aristotle. The logic literature has explored various ways of formally capturing them. The main approaches available include: Braham and Van Hees's game-theoretic framework in [1], stit logic accounts such as Canavotto [2], Ramírez-Abarca [7] and Duijf [3], multimodal accounts such as Glavaničová and Pascucci [5]. We present three scenarios where responsibility concerns despicable outcomes that materialized or could have materialized. None of the mentioned approaches can represent *all* three scenarios. In the second part of the work, the scenarios are formalized in a new framework of multimodal logic.

2 Scenarios

Scenario A: broken gun. *Martin aimed a gun at Benedict, thinking that he could kill him. Yet, when he pulled the trigger, the gun did not work. A police officer witnessed the scene and arrested Martin. Unknown to Martin, the gun lacked a component and was not capable of firing.*

Logical analysis of responsibility often focuses on the outcomes of an agent's actions, rather than on the actions themselves, and assumes that what an agent does is *causally effective* for an outcome; this is a trait of many works in stit logic [2, 7] and multimodal logic [5]. For instance, following [1], the agent's behaviour has to be a necessary element of a set of sufficient conditions for the outcome. Yet, in Scenario A, Martin's pulling the trigger had no causal connection with killing Benedict. We ascribe responsibility to Martin because he *believed* that pulling the trigger would *increase the likelihood* of killing Benedict. Thus, responsibility sometimes does not require any *objective* causal connection between a behaviour and an outcome. It is sufficient that there is a *reasonable subjective* connection, and it is not necessary that the outcome materializes (see Table 1). Only few approaches in the literature, such as [3], cover these cases.

Scenario B: child defense. *Anna's small daughter, Maria, is threatened by a thief who sneaked into their house. As an instinctive response to defend Maria, Anna grabs a knife and harms the thief. Later, a trial is held on this episode and no charges are brought against Anna.*

Despicable outcomes that are achieved are often sufficient for responsibility ascriptions in the logic literature [1, 3, 7]. In Scenario B, Anna freely and intentionally achieves an outcome that is problematic on its own: harming another person. Yet, she is not held legally (morally) responsible in the sense of being sanctionable (blameworthy). Indeed, we consider the context where the outcome occurred: Anna complied with a more important norm, namely protecting her child. Representing the *hierarchy of applicable norms* is often necessary to understand why responsibility is sometimes not ascribed despite the fact that all relevant criteria are met (Table 1). Again, only few approaches in the literature, such as [5], deal with this problem.

Scenario C: sabotaged project. *Immanuel wants his end-of-year project to be better than Ludwig's. Thus, he prepared an infected USB stick, which he plans to insert into Ludwig's laptop to prevent the latter from submitting his project. Yet, Immanuel accidentally exchanges his USB stick with Ludwig's and inserts the harmless USB stick into Ludwig's laptop. Later, Ludwig inserts the infected USB into his own laptop and is therefore unable to submit the project.*

In Scenario C, Immanuel has a reasonable belief that the action he intends to perform (inserting his own USB stick) will bring about a despicable outcome. However, he turns out to perform a different action (inserting Ludwig's USB stick). The latter action still has a causal impact on the outcome, allowing it to materialize, but only due to reasons that Immanuel had not considered (Ludwig will later insert the USB stick available to him, i.e. the infected one). Immanuel can be held responsible because he wanted to harm his classmate and took steps to achieve this result, even if the steps were not those he intended (Table 1). This scenario highlights that an agent can be held responsible even if they acted in a way that they had not regarded as a means to a reprehensible outcome, contrary to the analyses in [1, 7, 3].

Scenario	Freedom	Causation	Belief	Intented action	Outcome	Responsibility
A	✓		✓	✓		✓
B	✓	✓	✓	✓	✓	
C	✓	✓	✓		✓	✓

Table 1: Features of the three scenarios

3 Formal setting

We consider a language $\mathcal{L}$ with the following primitive symbols: a set of atomic sentence letters, labeled by ATOMS and whose members are denoted by P_1, P_2, P_3, etc.; the classical operators for material implication ($\rightarrow$) and negation ($\neg$); a set of norms, labeled by NORMS and whose members are denoted by n_1, n_2, n_3, etc.; a set of names for individual agents, labeled by AGENTS and whose members are denoted by i_1, i_2, i_3, etc; two unary operators of past necessity ($\blacksquare$) and future necessity ($\Box$); an indexed unary operator of obligation ($\mathcal{O}$); an indexed unary operator of intention/aim ($\mathcal{A}$); a binary relation of norm priority ($\blacktriangleleft$); an indexed binary operator of comparative possibility ($\preccurlyeq$); round brackets as auxiliary symbols.

Definition 1 (Formulas)**.** *χ is a formula of $\mathcal{L}$ precisely when it has one of the following forms, where $P \in \mathsf{ATOMS}$, $n, n' \in \mathsf{NORMS}$, $i \in \mathsf{AGENTS}$ and ϕ, ψ already count as formulas of $\mathcal{L}$:*

$$\chi ::= P \mid n \blacktriangleleft n' \mid \neg\phi \mid \phi \rightarrow \psi \mid \blacksquare\phi \mid \Box\phi \mid \mathcal{O}_i^n\phi \mid \mathcal{A}_i\phi \mid \phi \preccurlyeq_i \psi$$

The set of all formulas of $\mathcal{L}$ is denoted by FORM. We read formulas making use of non-classical operators as follows: $n \blacktriangleleft n'$ means that norm n is more authoritative than norm n'; $\blacksquare\phi$ that ϕ

holds throughout the past; $\Box\phi$ that ϕ holds throughout the future; $\mathcal{O}_i^n\phi$ that ϕ is obligatory for agent i in accordance with norm n; $\mathcal{A}_i\phi$ that agent i aims to obtain ϕ; $\phi \preccurlyeq_i \psi$ that agent i believes that ϕ is at least as likely as ψ. Other operators are defined as usual.

To interpret language $\mathcal{L}$, which contains an operator of comparative possibility ($\preccurlyeq$), we use a relational semantics endowed with a system of *nested spheres*, in the style of Lewis [6].

Definition 2 (Frames). *A relational frame for $\mathcal{L}$ is a tuple $\mathfrak{F} = \langle W, R, \$, f_o, h, f_a\rangle$ s.t.:*

- *W is a non-empty set of states;*
- *$R \subseteq (W \times W)$ is a relation of temporal precedence s.t.:*
  - *$R(w) = \{v \in W : (w, v) \in R\}$, the set of successors of w;*
  - *$R^{-1}(w) = \{v \in W : (v, w) \in R\}$, the set of predecessors of w.*
- *$\$: \mathsf{AGENTS} \times W \longrightarrow \wp(\wp(W))$ is a system of spheres s.t., for $i \in \mathsf{AGENTS}$ and $w \in W$:*
  - *$\$_i(w) \neq \varnothing$ and, for $S \in \$_i(w)$, $S \subseteq R(w)$ and $S \neq \varnothing$;*
  - *for $S, Z \in \$_i(w)$, $S \subseteq Z$ or $Z \subseteq S$.*
- *$f_o : \mathsf{NORMS} \times \mathsf{AGENTS} \times W \longrightarrow \wp(\mathsf{FORM})$ assigns a set of norm-relative obligations to agents at states;*
- *$h \subseteq \mathsf{NORMS} \times \mathsf{NORMS} \times W$ is a relation of priority between norms s.t.*
  - *for $n \in \mathsf{NORMS}$ and $w \in W$, $(n, n, w) \notin h$;*
  - *for $n, n', n'' \in \mathsf{NORMS}$ and $w \in W$, $(n, n', w), (n', n'', w) \in h$ implies $(n, n'', w) \in h$.*
- *$f_a : \mathsf{AGENTS} \times W \longrightarrow \wp(\mathsf{FORM})$ assigns a set of aims to agents at states.*

Definition 3 (Models). *A model for $\mathcal{L}$ is a tuple $\mathfrak{M} = \langle \mathfrak{F}, V\rangle$ where $\mathfrak{F}$ is a frame and:*

- *$V : W \longrightarrow \mathsf{ATOMS}$ is a valuation function.*

Definition 4 (Truth conditions). *Formulas of $\mathcal{L}$ are evaluated in accordance with the usual truth-conditions for classical operators and those listed below, where $[\phi] = \{w \in W : \mathfrak{M}, w \vDash \phi\}$:*

$\mathfrak{M}, w \vDash n \blacktriangleleft n'$ *iff* $(n, n', w) \in h$;
$\mathfrak{M}, w \vDash \blacksquare\phi$ *iff* $\mathfrak{M}, v \vDash \phi$ *for every* v *s.t.* $v \in R^{-1}(w)$;
$\mathfrak{M}, w \vDash \Box\phi$ *iff* $\mathfrak{M}, v \vDash \phi$ *for every* v *s.t.* $v \in R(w)$
$\mathfrak{M}, w \vDash \mathcal{O}_i^n\phi$ *iff* $\phi \in f_o(n, i, w)$
$\mathfrak{M}, w \vDash \mathcal{A}_i\phi$ *iff* $\phi \in f_a(i, w)$
$\mathfrak{M}, w \vDash \phi \preccurlyeq_i \psi$ *iff for every* $S \in \$_i(w)$, *if* $S \cap [\psi] \neq \varnothing$, *then* $S \cap [\phi] \neq \varnothing$.

4 Formal analysis of the scenarios

We use two defined operators, $\prec_i$ (taken from [6]) and $\Rightarrow_i$ (novel). The definition of $\prec_i$ is straightforward: $\phi \prec_i \psi =_{def} (\phi \preccurlyeq_i \psi) \wedge \neg(\psi \preccurlyeq_i \phi)$. A formula of the form $\phi \prec_i \psi$ indicates that agent i believes that ϕ is more likely than ψ. The definition of $\Rightarrow_i$ is more complex. Let $\theta_1 =_{def} (\phi \wedge \Diamond\psi)\prec_i(\phi \wedge \neg\Diamond\psi)$, $\theta_2 =_{def} (\phi \wedge \Diamond\psi)\prec_i(\neg\phi \wedge \Diamond\psi)$, $\theta_3 =_{def} (\neg\phi \wedge \neg\Diamond\psi)\prec_i(\neg\phi \wedge \Diamond\psi)$ and $\theta_4 =_{def} (\neg\phi \wedge \neg\Diamond\psi)\prec_i(\phi \wedge \neg\Diamond\psi)$. Then, $\phi \Rightarrow_i \psi =_{def} \theta_1 \wedge \theta_2 \wedge \theta_3 \wedge \theta_4$. An inspection of

the *definiens* reveals that a formula of the form $\phi \Rightarrow_i \psi$ indicates that agent i believes that ϕ *increases the likelihood* that ψ will occur.

We represent time in a scenario with *three temporal stages*, each corresponding to a class of states. Each state w belonging to the first two temporal stages is related *via* R to a range of possible successors belonging to the next stage; we only represent the successors that are associated with relevant possibilities for the scenario. The first stage always contains just one state, w_1. The second stage always contains four states, which are labelled as $w_{1.x}$, for $1 \leq x \leq 4$. The third stage always contains eight states, which are labelled as $w_{1.x.y}$ for $1 \leq x \leq 4$ and $1 \leq y \leq 2$. This notation allows us to keep track of alternative sequences of events. For instance, $w_{1.2.1}$ and $w_{1.2.2}$ are two successors of the same state ($w_{1.2}$), whereas $w_{1.2.1}$ and $w_{1.4.1}$ are not. Finally, each of the models used below contains information about aims and applicable norms (the latter are effective throughout the model).

Analysis of Scenario A. Let $\mathfrak{M}_1$ be the model in Figure 1, and let $w_{1.1.2}$ be the actual state of evaluation (marked in *orange*). Let m denote Martin, K stand for 'Benedict is killed', and P for 'Martin pulls the trigger'. The crucial state is w_1. At w_1, Martin aims to kill Benedict, $\mathfrak{M}_1, w_1 \vDash \mathcal{A}_m K$, and he is under an obligation not to do this, $\mathfrak{M}_1, w_1 \vDash \mathcal{O}^n_m \neg K$. Let arrows indicate the accessibility relation R, and let $\$_m(w_1) = \{S_1, S_2\}$, where $S_1 = \{w_{1.1}, w_{1.4}\}$ and $S_2 = S_1 \cup \{w_{1.2}, w_{1.3}\}$. *Green states* indicate possibilities that Martin *believes* to be more likely (i.e. S_1-states). We can see that $\mathfrak{M}_1, w_1 \vDash P \Rightarrow_m K$, which indicates that Martin believes that pulling the trigger increases the likelihood of killing Benedict. The outcome is not obtained, due to the gun being broken, and Martin's action is not causally successful at $w_{1.1.2}$. Yet, responsibility can be ascribed to Martin at $w_{1.1.2}$ due to his aim, belief and action.

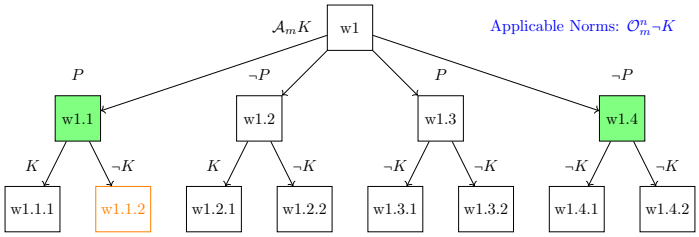

Figure 1: Model for Scenario A

Analysis of Scenario B. Let $\mathfrak{M}_2$ be the model in Figure 2 and $w_{1.1.1}$ be the actual state of evaluation. Let a denote Anna, H stand for 'the thief is harmed', S for 'Anna stabs the thief', and P for 'Maria is protected'. At w_1, Anna aims to protect her child and to harm the thief, $\mathfrak{M}_2, w_1 \vDash \mathcal{A}_a P$ and $\mathfrak{M}_2, w_1 \vDash \mathcal{A}_a H$. She is under two obligations: $\mathfrak{M}_2, w_1 \vDash \mathcal{O}^n_a \neg H$ and $\mathfrak{M}_2, w_1 \vDash \mathcal{O}^m_a P$, with m overriding n. Let $\$_a(w_1) = \{S_1, S_2\}$, where $S_1 = \{w_{1.1}, w_{1.4}\}$ and $S_2 = S_1 \cup \{w_{1.2}, w_{1.3}\}$. We can see that $\mathfrak{M}_2, w_1 \vDash S \Rightarrow_a H$ and $\mathfrak{M}_2, w_1 \vDash S \Rightarrow_a P$, indicating that Anna believes that stabbing the thief increases the likelihood of both harming him and protecting Maria. In this scenario, the fact that norm m overrides norm n plays a crucial role and we can ultimately exculpate Anna.

Analysis of Scenario C. Let $\mathfrak{M}_3$ be the model in Figure 3, and let $w_{1.2.1}$ be the actual state of evaluation. Let i denote Immanuel, S stand for 'Ludwig is sabotaged', and U for 'Immanuel uses his own USB stick'. At w_1, Immanuel aims to sabotage Ludwig, $\mathfrak{M}_3, w_1 \vDash \mathcal{A}_i S$, and is under an obligation not to do so, $\mathfrak{M}_3, w_1 \vDash \mathcal{O}^n_i \neg S$. Let $\$_i(w_1) = \{S_1, S_2\}$, where $S_1 = \{w_{1.1}, w_{1.4}\}$ and $S_2 = S_1 \cup \{w_{1.2}, w_{1.3}\}$. We have $\mathfrak{M}_3, w_1 \vDash U \Rightarrow_i S$, indicating that Immanuel believes that using his own USB stick increases the likelihood of sabotage. In this scenario, the outcome is

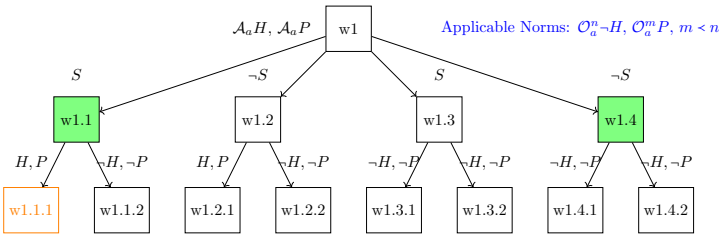

Figure 2: Model for Scenario B

obtained at $w_{1.2.1}$. Yet, this is not due to the action that Immanuel planned to perform in light of his belief (i.e. U). Responsibility can be ascribed to Immanuel because of his original aim.

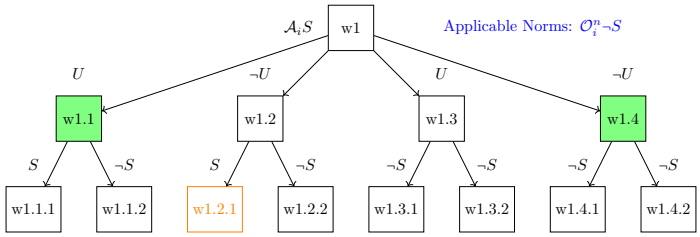

Figure 3: Model for Scenario C

Acknowledgments and contributions

The authors contributed equally to the article. The following statement applies to Matteo Pascucci: This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/I6499. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

References

[1] M. Braham and M. van Hees (2012). "An anatomy of moral responsibility." *Mind* 121(483): 601–34.

[2] I. Canavotto (2022). *Where Responsibility Takes You*. Cham: Springer.

[3] H. Duijf (2025). "A logical study of moral responsibility." *Erkenntnis* 90: 999–1040.

[4] J.M. Fischer and M. Ravizza (1998). *Responsibility and Control*. Cambridge: CUP.

[5] D. Glavaničová and M. Pascucci (2021). "Alternative semantics for normative reasoning with an application to responsibility and regret." *Logic and Logical Philosophy* 30: 653–679.

[6] D. Lewis (1973). *Counterfactuals*. Blackwell.

[7] I. Ramírez-Abarca (2022). *Logics of Responsibility*. PhD Thesis, University of Utrecht.