Responsibility in (Non-)Probabilistic STIT

David Streit[1]

University of Luxembourg
2, avenue de l'Université, L-4365 Esch-sur-Alzette
david.streit@uni.lu

Abstract. Attributing responsibility correctly when multiple agents are involved is important to both philosophy and artificial intelligence. STIT logics offer a compelling starting point for such a project. However, opposing proposals on how to do it have been put forward. In this paper, we investigate two such prominent proposals and argue that the apparent conflict between the two can be dissolved by looking at them through a probabilistic lens. To do so, we introduce a simple probabilistic STIT logic and two probabilistic responsibility predicates based on different background assumptions. We show that the (non-probabilistic) responsibility predicates discussed correspond exactly to the non-probabilistic fragment of the probabilistic ones. We show that we can generate counterexamples to each of the non-probabilistic proposals and that ultimately, such counterexamples cannot answer the question which one is preferable over the other.

Keywords: responsibility · stit · actual causation · blameworthiness.

1 Introduction

STIT logics cannot fully encode causal reasoning like Structural Causal Models can. Nevertheless, they allow for some counterfactual reasoning. "Were I to do (or had I done) action A, would it be the case that B?" is a question STIT can, and was designed to, answer. STIT also abstracts away from particular actions and their preconditions and instead focuses on outcomes. This seems to make it a good fit for a minimal logic of responsibility. We might be inclined to use it to answer questions of the type: "Something happened; who did it?" We will see that there are competing proposals on how to tackle this question. We will argue for two things. First, that a perceived disagreement between two recent ones, Naumov and Tao [9, 8] on the one hand and Sergot [10] on the other, is based on unexamined background assumptions and that the question of which is more appropriate ultimately has no answer. It depends. The way we will arrive at that conclusion is by looking at it through a probabilistic lens and by introducing two probabilistic versions of the proposals that will help put them in starker contrast. Second, we argue that ultimately, in cases where the outcome is "chancy", i. e. non-deterministic, non-probabilistic versions of a responsibility predicate in STIT are never fully satisfying. There are cases where the probabilistic responsibility predicates agree, both ascribe responsibility, but the non-probabilistic

ones do not. Before that, however, we will briefly give a gloss of what we mean, when we talk about responsibility.
For the kind of responsibility we are after to apply, traditionally, three features need to be present.[1] First, the actions of an agent must play a role in the occurrence of the outcome. Second, the actions of the agent were sufficiently free and third, there was a suitable alternative available to the agent that did not help produce the outcome (or did so to a lesser degree).
For present purposes, we ignore the second condition and focus on the first and third, as they are most naturally associated with STIT logics. We are additionally interested in a much more controversial phenomenon: group responsibility. Here we follow the somewhat crude approach in the literature and treat groups as arbitrary collections of agents whose actions are simply the simultaneous actions of its members. In treating groups like this, both the first and the third condition mentioned above makes sense if applied to the group level. Groups, in the sense of a collection of agents, can play a role in producing some outcome and groups, in the sense of the actions of their members, can have alternative actions available. We don't intend to commit to the position that responsibility proper can be attached to groups, but in the limited model that STIT offers, the responsibility we can ascribe to single agents can be ascribed to groups as well.

2 Responsibility as a Defined Formula

In this section, we will briefly discuss two different types of proposals of how to define responsibility as a defined predicate in STIT.[2] In the next section, we will show that counterexamples exist for both of these proposals. This is because, as we will show, both proposals turn out to correspond to different operators in a probabilistic version of STIT. We argue that the underlying intuition that one is responsible for an outcome if one could have made the outcome coming about less likely is what is behind the intuitions in play even in the non-probabilistic version. We further argue that there are two different interpretations of what that means, and that each interpretation produces one definition in probabilistic STIT. These two definitions in turn have as non-probabilistic correspondences exactly the two proposals discussed here. Thus, we argue that one should not prefer one over the other proposal at all as they rely on equally valid, but different, background assumptions.

The logic we will introduce below is a timeless STIT logic. It will consist of a set of worlds that are partitioned, for each agent, into equivalence classes. An

[1] While each feature is contested and disagreement persists with regard to the details, for present purposes, we follow Braham and van Hees [4] in treating this as a fair sketch of the default position. We also skip over the fact that in the context of STIT, what we call responsibility has been called actual causation or even blameworthiness. These semantic distinctions don't matter here.

[2] These are by no means the only approaches to responsibility (see Halpern [7]) and not even the only approaches to responsibility in STIT. However, we consider them prototypical for how responsibility has been approached in STIT.

equivalence class represents some nameless action an agent takes. Usually one is interested in evaluating from a specific world in a model. This world represents "the world as it is". It also gives us the action taken by every agent in that world via the equivalence classes the world is a part of. We say that an agent "sees to it that ϕ", iff the agent guarantees the outcome by her action. That is, if it is true in every world that is part of the same equivalence class as the world we are evaluating from. This is very similar to the "traditional" STIT logics including time, as explored in [2] by Belnap et al. in that it treats actions as equivalence classes of points of evaluation. We think the main lines of the arguments pursued here can equally be applied to a STIT logic with time, but chose to keep the base logic as simple as possible to avoid complications where none need be.

Then, let's define the timeless logic that allows for group actions as well. Let $\mathcal{P}$ be a countable set of atomic propositions, then the well-formed formulas are given by the following BNF, where $p \in \mathcal{P}$.

$$p \mid \Box\phi \mid [G]\phi \mid \neg\phi \mid \phi \to \phi$$

The Box operator stands for necessity, while the stit operator $[G]$ stands for *the group of agents G sees to it that*. Other propositional connectives are defined as usual.
Semantically,[3] the logic corresponds to the logic for *distributed knowledge* in epistemic logic. We use this logic as it has been explicitly endorsed in the literature[4] and because it is the simplest logic that allows us to achieve our purposes.

Definition 1. *A tuple $(W, \{R_\alpha\}_{\alpha \in Ag}, V)$ is called a model of STIT, where W is a non-empty set of worlds, R_α is an equivalence relation for every $\alpha \in Ag$, where Ag is a finite set of agents and $V : W \times P \to \{T, F\}$ is a propositional valuation function. For ease of exposition we further assume W to be finite, we leave it for future work to relax this condition.*

The accessibility relations can be lifted to the group level in a straightforward manner, allowing us to define the truth conditions.

$$R_G = \bigcap_{\alpha \in G} R_\alpha$$

Definition 2. *Given a model $M = (W, \{R_\alpha\}_{\alpha \in Ag}, V)$, and a world $w \in W$, we define when a formula ϕ is true at w, symbolically $M, w \models \phi$ inductively.*

[3] For reasons of space, we limit our discussion to the semantic part and only prove the most important statements explicitly in the appendix. The proofs for all other claims in this paper follow straightforwardly from the definitions and elementary semantic principles of modal logic.

[4] Naumov and Tao in [9, 8] and Sergot in [10] use it as a base logic. In a later paper, [8], Naumov and Tao add an S5 knowledge component. Semantically, however, this is inert and doesn't change the semantics of the responsibility operator. One additional assumption is *independence of agents* that is usually added in STIT (but consciously left out in Sergot [10]). We take no stance on the matter. Adding it or leaving it out leaves the main argument of this paper untouched.

- $M, w \models p$ *iff* $(w, p) = T$ *where* $p \in \mathcal{P}$.
- $M, w \models \neg\phi$ *iff* $M, w \not\models \phi$.
- $M, w \models \phi \rightarrow \psi$ *iff* $M, w \models \psi$ *or* $M, w \not\models \phi$.
- $M, w \models \Box\phi$ *iff for every* $v \in W$, $M, v \models \phi$.
- $M, w \models [G]\phi$ *iff for every* v, *if* wR_Gv *then* $M, v \models \phi$.

We introduce additionally the dual operator of $[G]$, $\langle G\rangle$. We define $\langle G\rangle\phi$ as $\neg[G]\neg\phi$. It is easy to verify that this corresponds to the semantic condition that:

$M, w \models \langle G\rangle\phi$ iff there exists a v, such that wR_Gv and $M, v \models \phi$

Intuitively, in a given world w, $\langle G\rangle\phi$ says that ϕ is a possible outcome given the actions of G in w. Another way of looking at it is saying that ϕ is a possible outcome when one holds fixed the actions of G. For the upcoming sections, the combination of the operators $\langle G\rangle$ and $[H]$ will become important. Here, too, we can find an intuitive meaning. $\langle G\rangle[H]\phi$ says, in effect, that H can guarantee that ϕ holds, regardless of (or because) the actions G does in w.

This simple logic allows us to write down three proposals[5] for a defined responsibility predicate in STIT. The first one we will simply call *active responsibility 1*.

Definition 3.

$$AR1(G)\phi \equiv [G]\phi \wedge \neg\Box\phi$$

The second one, to our knowledge first proposed by Bentzen in [3], but also recently by Naumov and Tao in [9, 8] we will call *passive responsibility 1*.

Definition 4 (Passive Responsibility 1).

$$PR1(G)\phi \equiv \phi \wedge \Diamond[G]\neg\phi$$

And lastly, proposed by Sergot in [10] as a response of perceived shortcomings of the previous definition, *passive responsibility 2*.

Definition 5 (Passive Responsibility 2).

$$PR2(G)\phi \equiv \phi \wedge \langle Ag \setminus G\rangle[Ag]\neg\phi$$

It is illustrative to show that all of these proposed definitions get simple cases right. Consider a case of underdetermination.

> **[Car]** Alice and Bob see a car parked on a hill that was left in neutral by the owner. If both of them push the car, they can push it downhill, causing the car to crash into a wall. Individually, Alice and Bob are too weak to do it. Alice and Bob both push and the car crashes. Alice and Bob, as a group, are responsible for the car crashing. (They might also be individually responsible, but for now we bracket that question).

[5] For each of these proposals, it is possible to add a minimality condition, demanding that the group of agents is the smallest group that is responsible. We ignore these throughout as they do not make the points presented more clear.

We can model this in a simple manner by stipulating a model with four worlds $W = \{w_1, w_2, w_3, w_4\}$. w_1 and w_2 correspond to the worlds where Alice pushes and w_3 and w_4 are the worlds where she does not. w_1 and w_3 are the worlds where Bob pushes. That is, $\{w_1, w_2\}$ and $\{w_3, w_4\}$ are the equivalence classes of R_{Alice}, while $\{w_1, w_3\}$ and $\{w_2, w_4\}$ are those of R_{Bob}. In order to describe the scenario faithfully, we require that V is such that $w_1 \models crash$ and $crash$ holds in no other world. We then get that $w_1 \models [\{Alice, Bob\}]crash \wedge \neg\Box crash$ satisfying the definition of active responsibility 1. But we also get $w_1 \models crash \wedge \Diamond[\{Alice, Bob\}]\neg crash$ or identically, $w_1 \models crash \wedge \langle\{Alice, Bob\} \setminus \{Alice, Bob\}\rangle \{Alice, Bob\}\neg crash$ and so passive responsibility 1 and 2 are true in w_1 as well. Before we ask ourselves if we can decide between the two, it is important to state that given the semantics we've introduced so far, all three definitions of responsibility are independent. We give one counterexample to the least obvious case in the appendix.

So how can we decide between the two? Sergot suggests in [10] that passive responsibility 2 is a more robust definition, and gives the following (slightly adapted) scenario to justify this position.

> [**Vase**] Alice and Bob can both carry a vase outside. If (at least) one of them does so and it rains the vase is ruined, if both leave the vase inside, or it does not rain, it is fine. It is a matter of chance (as far as the modeling of the scenario is concerned) whether it rains. Alice, but not Bob, takes the vase outside, and it does in fact rain, the vase is ruined. Alice (but not Bob) is responsible. She is not actively responsible by guaranteeing an avoidable outcome, but she is passively responsible by not preventing a preventable outcome.

We can model this as follows. Let $W = \{w_1, ..., w_8\}$ and $w_1, ..., w_4$ be the worlds where Alice puts the vase outside, and w_1, w_2, w_5, w_6 be the worlds where Bob puts the vase outside. We further assume that it is raining in the odd numbered worlds and doesn't rain in the even numbered worlds. The world corresponding to our scenario is then w_3. We know that $w_3 \models ruined$. Note, however, that it is not the case that $w_3 \models \Diamond[\{Alice\}]\neg ruined$ as Alice can never guarantee that Bob doesn't take the vase outside (and it rains). Thus, in w_3 we do not get passive responsibility 1. As Alice does not guarantee the outcome (because it could rain or not), she is also not responsible according to active responsibility 1.
We do however get that $w_3 \models \langle\{Alice, Bob\} \setminus \{Alice\}\rangle[\{Alice, Bob\}]\neg ruined$ as, given Bob's not carrying the vase outside, Alice can guarantee that the vase stays inside and so in w_3, passive responsibility 2 does hold.
As the verdict that Alice is the solely responsible actor in this scenario is highly intuitive, this is a strike against passive responsibility 1. So should one follow Sergot in claiming that what best determines responsibility is what one could prevent if one holds fix other people's actions? This would be too hasty. Consider the following case

> [**Blanks**] Alice and Bob are made to shoot an innocent person. Each can load their gun with either a blank or a (normal) bullet at their behest. If

> hit by a blank, the victim will be left unharmed (by the blank). If hit by a single bullet, the victim might die or might live, if the victim is hit by two bullets, she will die. All of this is known to Alice and Bob, who load their weapons with bullets and shoot. The victim dies. Here both Alice and Bob are *individually* (at least partially) responsible for the death.

If we model this in the style of the earlier examples, Alice and Bob are not individually responsible under passive responsibility 1 or 2, neither are they responsible according to active responsibility 1, even though we clearly want them to be responsible in *some* sense. What can one do? One option is to try to find yet another definition for responsibility that can capture the intuitions in this example. We choose a different avenue. We will show that if one looks at the definitions and examples through a probabilistic lens, one can explain these problems. In fact, in the end, we will find that a (non-probabilistic) definition that captures the notion of responsibility in this example will be a byproduct of this line of reasoning.

3 Probabilistic STIT

We claim that the intuitions generated in all the preceding examples can be best explained by appealing to the following principle: An agent or a group of agents is responsible for an outcome iff they had an action available to them that made the outcome less likely.[6] We will see that there are two interpretations one can give that closely correspond to passive responsibility 1 and 2 and that in a probabilistic setting both definitions are generated by different background assumptions. We do this by introducing a simple probabilistic version of the logic introduced previously, and show that the probabilistic version explains the appeal and shortcomings of the definitions of responsibility introduced before. The probabilistic version is build in the style of Fagin et al. in [6] and van Eijck and Schwarzentruber in [11], while keeping it as simple as possible.[7] Like before, let $\mathcal{P}$ be a countable set of atomic propositions, $p \in \mathcal{P}$, and $k \in \mathbb{Q} \cap [0,1]$. Then the well-formed formulas are given by the following BNF:

$$p \mid \Box\phi \mid [G]^{p=k}\phi \mid \neg\phi \mid \phi \to \phi \mid k < k$$

Additional terms can be defined from these, for example $k \geq k'$ as $\neg(k < k')$.

Definition 6. *A model for probabilistic STIT is a tuple* $(W, \{R_\alpha\}_{\alpha \in Ag}, P, V)$ *where* $(W, \{R_\alpha\}_{\alpha \in Ag}, V)$ *is a STIT model and* $P : W \to \mathbb{Q}_+ \setminus \{0\}$. *Further, we demand that* $\sum_{w \in W} P(w) \neq \infty$.

[6] We bracket the question of how exactly to interpret probability here. For now, treat it as an objective probability of an outcome occurring.

[7] To have an axiomatic system, it is fruitful to stay closer to [6]. However, as this is not the focus here, an anonymous reviewer rightly suggested keepingthe semantics simple.

Like before, the accessibility relation for groups is lifted from the accessibility relations of its members and the Box operator is defined as the universal modality.

To keep the semantics tractable, we introduce the following definition. Let $|w|^{\phi}_{G}$ be the set $\{v : wR_Gv \text{ and } M, v \models \phi\}$ This allows us to define truth in a world in the probabilistic model analogously.

Definition 7. *Given a probabilistic STIT model $M = (W, \{R_\alpha\}_{\alpha \in Ag}, P, V)$, and a world $w \in W$, we define when a formula ϕ is true at w, symbolically $M, w \models \phi$ inductively. The conditions for propositional atoms, propositional connectives and the box operator are omitted here. They are identical to the non-probabilistic version introduced above.*

- $M, w \models t < t'$ *iff* $t < t'$
- $M, w \models [G]^{p=t}\phi$ *iff* $\frac{\sum_{v \in |w|^{\phi}_{G}} P(v)}{\sum_{v \in |w|^{\top}_{G}} P(v)} = t$

The first thing we need to make sure is that the sums are always well-defined and that the quotient is indeed always a rational number between 0 and 1. By definition, $\sum_{u \in W} P(u) < \infty$ and for every w, G and ϕ, $|w|^{\phi}_{G} \subseteq W$ and so too $\sum_{v \in |w|^{\phi}_{G}} P(v) < \infty$. The quotient is then defined, iff $\sum_{v \in |w|^{\top}_{G}} P(v) \neq 0$. Note that since $|w|^{\top}_{G}$ is an equivalence class, $w \in |w|^{\top}_{G}$. As, by definition, $P(w) > 0$, $\sum_{v \in |w|^{\top}_{G}} P(v) > P(w) > 0$. So the quotient is always defined. All that remains to be checked is if it is a rational number between 0 and 1. This follows immediately as $|w|^{\top}_{G} \supseteq |w|^{\phi}_{G}$ and so the numerator of the quotient can only be at most as large as the denominator, which guarantees that it stays between 0 and 1.

But how are we to interpret $P(w)$? Given some action by a group G, such that w and v are both possible outcomes of that action, i. e. they are in the same equivalence class for that action, $P(w)$ and $P(v)$ express the relative likelihood, given the actions of G, of w or v being the world that is the outcome of the action. So if $P(v) = 3$ and $P(w) = 6$, then it is twice as likely that we end up in world w than it is that we end up in v prior to taking the action. What $M, w \models [G]^{p=t}\phi$ says is then that given the action(s) the group G performs in w, the chance that, given the actions of G, we end up in a ϕ-world is (exactly) t.

It is possible to introduce the non-probabilistic group STIT operator in the probabilistic version using the same definition as before.

$M, w \models [G]\phi$ iff for every v, if wR_Gv then $M, v \models \phi$.

As we don't allow worlds with value zero, it is easy to verify that the semantic conditions of $[G]$ and $[G]^{p=1}$ are the same. Hence, we will use them interchangeably in the probabilistic logic. We will also continue to use $\langle G \rangle$ using either the semantic definition of the non-probabilistic logic or (equivalently) defined as $\neg[G]^{p=1}\neg\phi$.

Next we'll define and only later discuss probabilistic versions of responsibility that resemble closely passive responsibility 1 and 2. We will call these p-responsibility 1 and p-responsibility 2 respectively.

Definition 8 (P-Responsibility 1).

$$ProbR1(G)\phi \equiv \phi \wedge [G]^{p=k}\phi \wedge \Diamond[G]^{p=c}\phi \wedge k > c$$

Definition 9 (P-Responsibility 2).

$$ProbR2(G)\phi \equiv \phi \wedge [Ag]^{p=k}\phi \wedge \langle Ag \setminus G\rangle[Ag]^{p=c} \wedge k > c$$

These two definitions express different notions of responsibility. The first one claims that a group is responsible only if it has an alternative option available that reduces the likelihood of the outcome regardless of the other agent's actions while the second one takes them as given and only ascribes responsibility if, given those actions taken as fixed, the group can reduce the likelihood of the outcome.

A quick note is in order. As there are infinitely many values k and c can take in the definitions, we might be tempted to add an existential quantifier to our logical language and treat the definitions as an existentially quantified formula. However, this adds unwanted complexity. There are two ways around this without complicating the logic further. Either, one treats the definition as a schema that is shorthand for a countable amount of definitions, or one restricts the possible values k and c can take to be finite. Then the definition can be seen as a (finite) disjunction of formulas, one for each combination of values for k and c. We will skip these complications here.

Depending on circumstances, both definitions can be plausible. For example, if one assumes that the group of agents knows what other agents will do, then clearly the second one is better. If one assumes that they do not, then the first one seems more appropriate.[8] So given two different background assumptions on what one wants to model, the definitions diverge.

We can also see that passive responsibility 1 entails p-responsibility 1 (but not 2) and passive responsibility 2 entails p-responsibility 2 (but not 1). Proofs of this entailment can be found in more detail in the appendix. In fact, all three definitions of (non-probabilistic) responsibility are obtainable by setting the probabilities in the definitions to 0 or 1 (in effect, replacing them by their non-probabilistic fragment).

As we have seen, we can replace a probabilistic operator $[G]^{p=t}$ by a non-probabilistic operator if, and only if, $t = 0$ or $t = 1$. As both formulas of p-responsibility contain two probabilistic operators, we have two points of attack. It will turn out, that for both probabilistic definitions of responsibility, there are exactly two ways to turn the operators into non-probabilistic ones.

Let us go through them one by one. For p-responsibility 1, defined as $\phi \wedge [G]^{p=k}\phi \wedge \Diamond[G]^{p=c}\phi \wedge k > c$, one has only two possibilities. Either one sets $k = 1$ and c some (arbitrary) value less than 1 or one sets $c = 0$ and k some (arbitrary) value larger than 0. In the former case, when $k = 1$, this implies

[8] We don't claim that the epistemic reading is the only one that makes a difference between the two, we only think it presents a clear case where the two definitions come apart and that it makes it clear that these two express genuinely different notions.

(non-probabilistic) active responsibility 1. In the latter case, when $c = 0$, this implies passive responsibility 1. Proofs of these statements can be found in the appendix.

What does this tell us? Active and passive responsibility 1 are special cases of the probabilistic operator, they are derivable from the limit cases where one of the probabilistic operators is respectively assigned either a probability of 1 or 0.

We can find a similar fact for p-responsibility 2. Once again, in the definition $\phi \wedge [Ag]^{p=k}\phi \wedge \langle Ag \setminus G\rangle[Ag]^{p=c} \wedge k > c$, we can either set $k = 1$ or $c = 0$. Like before, if one sets $c = 0$, we obtain a form of passive responsibility, only this time we are left with a formula equivalent to passive responsibility 2.

But what if one sets $k = 1$? If we simplify the formula a bit, we are left with $[Ag]\phi \wedge \langle Ag \setminus G\rangle\neg[Ag]\phi$. This is the active responsibility fragment of p-responsibility 2 which we will call active responsibility 2. To our knowledge, it has not been suggested anywhere in the literature.

That it has not been is a curious fact. The principle says that one is responsible if, given the actions of the other agents, one guarantees an outcome but, given those actions, one didn't need to. This is exactly the principle we needed for the *Blanks* example. According to active responsibility 2, Alice (and Bob) are responsible as each has an option that, given the other person's actions, does not make the death of the victim guaranteed.

Why not stop here and say that there simply are two sets of different but equally valid principles of responsibility. The reason is that in the *Vase* example, responsibility ascription is easy. It doesn't seem to matter if Alice knows what Bob is going to do or not, if she puts the vase outside, and it rains, she is responsible. As both active and passive responsibility 1 cannot, but passive responsibility 2 can accommodate this, is this not a strike against p-responsibility 1 as well?

If we adapt the model corresponding to *Vase* to be probabilistic,[9] we see that in fact, p-responsibility 1 returns the correct result verdict here as well. This is why the vase example appears to be so compelling. Both probabilistic versions of responsibility agree on the verdict, but the non-probabilistic fragments do not. We claim that this is not a strike against passive responsibility 1. It is almost accidental that we can represent this case faithfully non-probabilistically under one (probabilistic) interpretation but not under another.

We have seen, if one accepts a probabilistic viewpoint, there are two genuinely different principles of responsibility one can define depending on whether one considers the actions of other agents as fixed or not. These principles in turn generate each one passive and one active non-probabilistic responsibility predicate. However, even if the model is non-probabilistic, none of them can accommodate all cases of chancy causation. In fact, focusing on counterexamples is the wrong approach, counterexamples exist to each non-probabilistic predicate. Thus, instead of trading counterexamples, an analysis of responsibility in STIT should fall back on the probabilistic understanding that explains the different

[9] We set $P(w_i) = q_i$, with some arbitrary positive q_i for each $w_i \in W$.

meanings of the principles in question, choosing the appropriate definition based on what situation one is modeling.

4 Related and Future Work

We restrict ourselves a very brief discussion of closely related treatment of responsibility in game theoretic or STIT settings, while acknowledging that there are other suggestions of to make sense of responsibility. Most importantly, we will not focus on ways to distribute responsibility for an outcome to individuals, as suggested by Yazdanpanah et al. in [12] and Chockler and Halpern in [5]. Instead, we will focus our attention on papers that, like us, aim to answer the yes/no question is a (group of) agent(s) responsible?
In [4], Braham and van Hees give what they call an anatomy of moral responsibility in game theoretical terms. There are two main differences to highlight. First, they do not attempt to lift responsibility to the group level and second, they filter possible actions through a solution concept. We don't think this is a helpful way to look at responsibility generally. We think that if they attempted to lift their proposal to the group level, they would have to make exactly the choice we present here. Treat other agent's actions as fixed or not.
The second interesting suggestion is Baltag et al. in [1]. Here, responsibility is defined in STIT by adding the notion of opposing actions. We believe that this does not help us solve the problem. In "chancy" cases, including the scenarios above, unlike the probabilistic version introduced here, any plausible formalization in Baltag et al.'s system undergenerates responsibility ascriptions which we take to point towards the fact that here as well non-probabilistic STIT cannot account for all cases.
In the future, we aim to fully axiomatize the probabilistic logic and extend it to more expressive STIT frameworks like XSTIT. Additionally, it is useful to add a more explicit treatment of epistemic states, both through the addition of subjective probabilities and through the addition of knowledge and belief operators, allowing for a more full-fledged treatment of the subjective part of responsibility. Finally, as blameworthiness and praiseworthiness are intimately connected but not the same as responsibility, it is worth investigating how the addition of deontic components to the logic can make sense of this connection.

Appendix

Theorem 1. *Passive responsibility 2 does not entail passive responsibility 1.*

Proof. Consider as a counterexample a model M with three worlds w_1, w_2, w_3 and two agents A and B with $w_1 R_A w_2$ and $w_2 R_B w_3$. Further, assume that $M, w_1 \models p$, $M, w_2 \models p$ and $w_3 \models \neg p$. Then, $w_1 \models [A]p \wedge \Diamond[A]\neg p$, satisfying passive responsibility 1. However, as w_1 is the only world accessible by $R_{Ag\setminus A} = R_B$ is w_1 itself, and $[Ag]\neg p$ does not hold in w_1, $M, w_1 \not\models \langle Ag - A\rangle[Ag]\neg p$ and so passive responsibility 2 does not hold in w_1.

Definition 10 (Probabilistic model based on a non-probabilistic model). *We say a probabilistic model $M' = (W, \{R_\alpha\}_{\alpha \in Ag}, P, V)$ is based on a non-probabilistic model M iff $M = (W, \{R_\alpha\}_{\alpha \in Ag}, V)$.*

Theorem 2. *Passive responsibility 1 entails p-responsibility 1. We need to make this more precise. If for some non-probabilistic model $M = (W, \{R_\alpha\}_{\alpha \in Ag}, V)$ and $w \in W$, $M, w \models PR1(G)\phi$, then, for every probabilistic model M' based on M, $M', w \models ProbR1(G)\phi$.*

Proof. By assumption, $PR1(G)\phi$ is true in w in M. So $M, w \models \phi \wedge \Diamond[G]\neg\phi$. As the only difference between the probabilistic model M' and M is the existence of a function P that assigns weights to worlds and truth in w is defined analogously, for non-probabilistic operators, the truth conditions are the same in both models. We've also seen that in the probabilistic model, $[G]$ corresponds to $[G]^{p=1}$, we can translate the formula to the probabilistic logic while preserving truth in w. And so we know that $M', w \models \phi \wedge \Diamond[G]^{p=1}\neg\phi$. Specifically, because there are no worlds with weight 0, there is a (non-empty) equivalence class for G, such that $\neg\phi$ holds in every world in the class. This means that in this class, no world is a ϕ world. And so $M', w \models \Diamond[G]^{p=0}\phi$. But we also know that $M', w \models \phi$. We know that $P(w) > 0$ and so, since w is in the equivalence class of w for G, $M', w \models [G]^{p=k}\phi$ where $k \geq P(w) > 0$. And so $M', w \models \phi \wedge [G]^{p=k}\phi \wedge \Diamond[G]^{p=0}\phi \wedge k > 0$. And so, by definition, $ProbR1(G)\phi$ holds in w.

Theorem 3. *Active responsibility 1 entails p-responsibility 1. Let M be a non-probabilistic model with $M = (W, \{R_\alpha\}_{\alpha \in Ag}, V)$ and $w \in W$, $M, w \models AR1(G)\phi$, then, for every probabilistic model M' based on M, $M', w \models ProbR1(G)\phi$.*

Proof. By assumption, $[G]\phi \wedge \neg\Box\phi$ is true in w for M. By analogous reasoning as in the last proof, this means that $M', w \models [G]^{p=1}\phi \wedge \neg\Box\phi$. This immediately implies that $M', w \models \phi$. As $M', w \models \neg\Box\phi$, there is some world v, such that $M', v \models \neg\phi$. This means that in the equivalence class of v for G, the sum of the weights of all worlds in the equivalence class is larger than the worlds of all ϕ worlds in the class, as v has non-zero weight. But this means, by definition, that $M', v \models [G]^{p=k}\phi$ for some $k < 1$, and so $M', w \models \Diamond[G]^{p=k}\phi$. Again, this is all we need. We have, $M', w \models \phi \wedge [G]^{p=1}\phi \wedge \Diamond[G]^{p=k}\phi \wedge 1 > k$ and so $M', w \models ProbR(G)\phi$.

Theorem 4. *Passive responsibility 2 entails p-responsibility 2. Let M be a non-probabilistic model with $M = (W, \{R_\alpha\}_{\alpha \in Ag}, V)$ and $w \in W$, $M, w \models PR2(G)\phi$, then, for every probabilistic model M' based on M, $M', w \models ProbR2(G)\phi$.*

Proof. By assumption, $\phi \wedge \langle Ag \setminus G\rangle[Ag]\neg\phi$ is true in w for M. Just like before, this implies that $M', w \models \phi \wedge \langle Ag \setminus G\rangle[Ag]^{p=0}\phi$. We know that the ϕ world w is in the equivalence class of w for Ag with $P(w) > 0$, and so $M', w \models [Ag]^{p=k}\phi$ with $k > 0$. With this we assembled all the necessary elements, since $M', w \models \phi \wedge [Ag]^{p=k}\phi \wedge [Ag]^{p=0} \wedge k > 0$ and so $M', w \models ProbR2(G)\phi$.

Theorem 5. *$\phi \wedge [G]^{p=1}\phi \wedge \Diamond[G]^{p=c}\phi \wedge 1 > c$ implies active responsibility 1.*

Proof. Let M be an arbitrary model and w a world of that model, such that $M, w \models \phi \wedge [G]^{p=1}\phi \wedge \Diamond[G]^{p=c}\phi \wedge 1 > c$. Then, $M, w \models [G]\phi$ since $M, w \models [G]^{p=1}\phi$. As $c < 1$, there is a world v, such that $M, v \models [G]^{p=c}$ and $v \models \neg\phi$. Hence $M, v \models \neg[G]\phi$ and hence $w \models \Diamond\neg[G]\phi$. This means that $M, w \models [G]\phi \wedge \Diamond\neg[G]\phi$. Which in turn means, by definition $M, w \models AR1(G)\phi$.

Theorem 6. $\phi \wedge [G]^{p=k}\phi \wedge \Diamond[G]^{p=0}\phi \wedge k > 0$ *implies passive responsibility 1.*

Proof. Let M be an arbitrary model and w a world of that model, such that $M, w \models \phi \wedge [G]^{p=k}\phi \wedge \Diamond[G]^{p=0}\phi \wedge k > 0$. Then, as $M, w \models \Diamond[G]^{p=0}\phi$, $M, w \models \Diamond[G]\neg\phi$ and so $M, w \models \phi \wedge \Diamond[G]\neg\phi$. But this is just the definition of passive responsibility 1.

Theorem 7. $\phi \wedge [Ag]^{p=k}\phi \wedge \langle Ag \backslash G\rangle [Ag]^{p=0} \wedge k > 0$ *implies passive responsibility 2.*

Proof. The proof follows the same lines as the preceding ones.

References

1. Baltag, A., Canavotto, I., Smets, S.: Causal agency and responsibility: a refinement of stit logic. In: Logic in High Definition, pp. 149–176. Springer (2021)
2. Belnap, N., Perloff, M., Xu, M.: Facing the future: agents and choices in our indeterminist world. Oxford University Press (2001)
3. Bentzen, M.M.: Stit, Iit, and Deontic Logic for Action Types. Ph.D. thesis, University of Amsterdam (2012)
4. Braham, M., Van Hees, M.: An anatomy of moral responsibility. Mind **121**(483), 601–634 (2012)
5. Chockler, H., Halpern, J.Y.: Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research **22**, 93–115 (2004)
6. Fagin, R., Halpern, J.Y., Megiddo, N.: A logic for reasoning about probabilities. Information and computation **87**(1-2), 78–128 (1990)
7. Halpern, J.Y.: Cause, responsibility and blame: a structural-model approach. Law, Probability and Risk **14**(2), 91–118 (01 2015)
8. Naumov, P., Tao, J.: An epistemic logic of blameworthiness. Artificial Intelligence **283**, 103269 (2020)
9. Naumov, P., Tao, J.: Two forms of responsibility in strategic games. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (2021)
10. Sergot, M.: Actual cause and chancy causation in: A preliminary account. In: Agency, Norms, Inquiry, and Artifacts: Essays in Honor of Risto Hilpinen, pp. 21–42. Springer (2022)
11. Van Eijck, J., Schwarzentruber, F.: Epistemic probability logic simplified. In: Advances in modal logic (2014)
12. Yazdanpanah, V., Dastani, M., Alechina, N., Logan, B., Jamroga, W.: Strategic responsibility under imperfect information. In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems AAMAS 2019. pp. 592–600. IFAAMAS (2019)