From Responsibility, via Indifference, to Recklessness

[INITIAL STEPS AND FORMALISATIONS]

Michael Fisher
Department of Computer Science, University of Manchester*
michael.fisher@manchester.ac.uk

The notion of 'responsibility' as a higher-level construct that dynamically impacts each agent's goals, priorities and actions is very appealing, especially as humans regularly use such concepts in everyday reasoning. Our aim is to utilise 'responsibility' to drive proactive *computational* agent behaviour and, importantly, to highlight when an agent need not do anything as well as when it should.

In this work, we look at formalising responsibility, and especially how the concept of responsibility leads to goals or actions within our agents. We are also interested in hierarchies of responsibility. For example, even though responsible for some aspect our agent might decide to do nothing if it believes some other agent is *more* responsible. We are also interested in the converse of responsibility – an agent *not* being responsible – and want to also use this to drive agent behaviour. In particular, there may be different varieties of this "lack of responsibility" – not just *irresponsibility* but *recklessness* and even *maliciousness* that we also aim to formalise.

1 Introduction

1.1 Why Responsibility?

Responsibility is important in driving much human activity, and we contend that the concept can also be useful in autonomous systems. At the core of autonomous systems is the concept of autonomous decision-making, the system being able to (and often needing to) make its own decisions without human oversight. There are many mechanisms for representing and assessing autonomous decision-making and we aim to show that high-level (and abstracted) concepts of "responsibility" might be useful in moving towards decisions, intentions, and actions.

The concept of responsibility is typically split into *backward* and *forward* versions. Backward responsibility, very close to concepts of accountability and blame, concerns what happened and who was responsible for it happening. However, we will focus on *forward responsibility*, or *prospective responsibility*. This impacts what we (or our autonomous systems) will do, and why, and so is a core driver in both actions and choices in the future.

A further element, one that will be useful in embodied systems such as robots, is the use of responsibility to lead the system to choose *not* to do something, even when it could. This is an important element of human decision-making and will help our autonomous systems move on from being purely reactive. In essence, while the general aim of responsibilities is to drive goals/intentions and then actions, we will typically choose inaction in one of three cases:

*Thanks to UKRI funding for the *Computational Agent Responsibility* project (EP/W01081X) and the Royal Academy of Engineering for their *Chair in Emerging Technologies* funding.

Submitted to:
FEAR 2024

1. if it's not our responsibility;
2. if it is our responsibility, but we cannot do anything about it; or
3. if it is our responsibility, but we believe we are not needed.

This final item comes in to play if we believe there are enough other agents, more responsible and capable, so that we do not need to act. We wish to have similar reasoning within our autonomous systems.

Example: Our hospital robot might have responsibilities for cleaning and tidying, for moving items and people around the hospital, but is also responsible for respecting patients' autonomy and wishes. The robot might also have responsibilities concerning reporting health issues and privacy, with obvious impact upon trust and safety. If, in its tidying duties, the robot detects a fire then the robot's responsibility for patient safety should generally override issues around human autonomy and the robot should immediately ensure the person is safely out of the hospital. And, while the robot might have a general responsibility for patient well-being, once a doctor or nurse is present, the robot reasons that it does not need to tend to the patient's needs as these "more responsible" humans will deal with this while present.

Once we can formally define "responsibility" then we can also examine its negation or the idea of a "lack of responsibility". We will explore variations on this theme later in the paper, defining terms such as "indifference" and even "recklessness".

A final variation is that while we predominantly consider "responsibility for" some issue, we will also briefly involve the separate, but related, concept of "responsibility to" other entities. This will be useful for delegation/responsibility hierarchies within groups or teams.

1.2 Why Agents?

We are concerned with analysing the concept of responsibility, but particularly if, and how, it might be used in practical, embodied systems. In particular, we are engaged in building autonomous systems that are *trustworthy* and *resilient*. Within these systems we focus on the core (autonomous) decision-making, especially formal verification of the way that decisions are made [14]. In this, we construct autonomous systems around cognitive/rational/intelligent *agents* [35] and specifically BDI Agents [29]. These agents expose the motivations/reasons behind decisions and the details of the decision-making processes and once we embed such an agent in our practical system as the core decision-making component, we are able to develop more transparent, verifiable, and reliable autonomous systems [16]. Specifically, such an agent giving the core decision-making capabilities within our autonomous systems ensures

- *transparency* – of behaviour, of intention, etc, so we can see what this agent will do (and why) [34]
- *verifiability* – again, of intention or of behaviour, for example proving that the agent always makes the 'right' decisions [14]
- *explainability* – based on transparency and, once verified, will truthfully and appropriately answer questions about its behaviour [19]

Consequently, we take an agent-based approach [31] as our basis, building the concept of "responsibility" on top of this.

The basic idea of responsibility is closely linked to that of a "role". While there has been vast literature on roles and role hierarchies in agent-based systems [4,9,27], we do not explicitly use this since we want to construct responsibility structures/hierarchies on-the-fly. We assume an open and distributed agent system with no centralised organisation, for example no prescribed roles or set role hierarchies. Clearly, the two approaches intersect when used in essentially fully understood environments, but we want to explore the individual concept of responsibility and how it affects agent action, irrespective of any larger context. One way to see this as slightly different is to view our approach as a *bottom-up* construction of dynamic responsibilities, rather than the *top-down* prescription of role hierarchies.

1.3 Why Formalise?

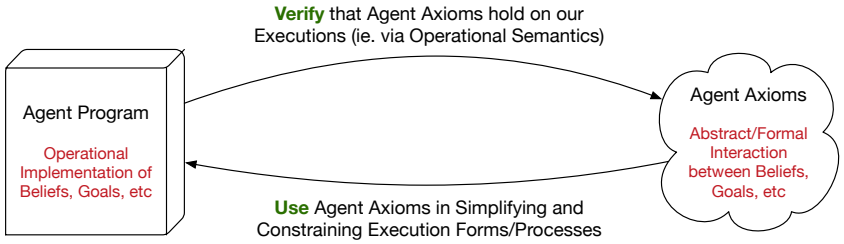
In the late 1980s and early 1990s, many papers providing logical foundations of agents were published. Often based on BDI but also Knowledge, Obligations, Actions, etc., a cottage industry built up in papers that examined the interactions between these different concepts, especially when described in some logical form. The use of modal logics, temporal logics, dynamic logics, situation calculus, etc, was common. Much of this was stimulated by Bratman's book [8], with leading researchers such as Cohen & Levesque [10], Konolige [20], Moore [24], Rao & Georgeff [30], etc. Many papers looked at interactions between (usually BDI) modalities and what these might represent (at least in principle), e.g[1]:

$$\mathsf{Intends}(\varphi) \wedge \Box\mathsf{Believes}(\Box(\varphi \Rightarrow \psi)) \quad \Rightarrow \quad \mathsf{Intends}(\psi)$$

or

$$\mathsf{Intends}(\varphi) \Rightarrow \neg\mathsf{Believes}(\Box\neg\varphi)$$

These sorts of developments are based on the close relationship between (agent) axioms and (agent) programs:



It is essentially this approach that we intend to follow, providing formalisations of "responsibility" and axiomatisations of how responsibility can lead to goals/intentions and actions in our agents.

In the next section we will review some of the existing works on Logics of Responsibility, but note that a distinguishing feature of our work is that we represent "responsibility" as a basic concept, not as one derived from other concepts, such as *obligations*. Our agents have responsibilities and these will drive the agent's action (or inaction).

[1] Here, $\Box$ is a linear temporal logic operator meaning "always in the future".

2 Previous Work in Related Areas

Below we highlight some of the relevant previous work formalising aspects of responsibility.

De Lima, Royakkers and Dignum [21]: In this paper, entitled *"A Logic for Reasoning about Responsibility"*, the authors describe both forward-responsibility and backward-responsibility but do this as an interaction between *knowledge* and *time* with additional elements of both ATL[2] and *obligation* included. A summary of their definition is:

> *agent i is forward-looking responsible for φ if, and only if, any state from which i does not have the power to ensure φ is a violation state*

Essentially, forward-responsibility is here closely linked to the agent's ability to bring φ about.

For backward-looking responsibility

> *agent i is backward-looking responsible for φ after the occurrence of event $\delta|_i$ if, and only if, i is responsible for $\neg\varphi$ and $\delta|_i$ is irresponsible for i.*

This, quite strong, constraint requires detailed/certain knowledge about the group of agents (where $\delta|_i$ describes what some group of agents can do) that can (or did) achieve some state.

Royakkers and Hughes [32]: Similar to, and involving a co-author from, the de Lima, Royakkers and Dignum article, Royakkers and Hughes extend the logic further, incorporating elements from Dynamic Epistemic Logic and Logics of Group Knowledge. As in the de Lima, Royakkers and Dignum article (though they claim, simplified), *forward-responsibility* is defined as a form of *obligation* while *accountability* is seen as a combination of *responsibility* and *causality*.

Glavaničová and Pascucci [18]: Though predominantly about *norms* and *accountability*, this paper includes a little about *prospective responsibility*: "responsibility for something that should obtain either now or in the future, according to some norm currently in effect". The authors define "prospective responsibility" essentially in deontic terms: an agent with prospective responsibility has a certain obligation(s) towards the present or the future. An attribution of prospective responsibility may also concern a sequence of states to be achieved and duties of other agents.

Lorini and Sartor [22, 23]: Both these papers are centred on (temporal) STIT, the logic of "Seeing To It That" [5, 6], and *knowledge*. Their focus is described as:

> *"we precisely characterise a notion of influence-based responsibility, namely, a responsibility that depends on the fact that an agent causes a primary violation by another agent"*

Rather than direct responsibility, this approach represents a form of *secondary*, or *indirect*, responsibility.

[2] Essentially, a Temporal Logic describing group capabilities and strategies [1].

Giordani [17]: Again based on action logics, knowledge and STIT, the paper states:

> *"The main advantage of the present system lies in the possibility of analysing the fact that an agent brings about a certain state of affairs in two distinct components: the fact that the agent performs a specific basic action and the fact that a state of affairs is a consequence of the performed action. This kind of analysis allows us to introduce a novel account of the notions of epistemic ability and knowingly doing and a comprehensive conceptual framework for classifying different levels of responsibility."*

Particularly relevant for *accountability* and *backwards-responsibility* for actions.

Yazdanpanah, et al [36]: Using a strategic approach again, this article concentrates mainly on backwards-responsibility. Responsibility is ascribed to runs through concurrent game structures and a verification route is provided via ATL (but only for the backwards-responsibility). As stated in the paper, a group of agents, Λ, is responsible for φ over a history σ if (1) φ occurs in σ, (2) the agents in Λ can together achieve φ, and (3) removal of any of the agents from Λ stops us being able to achieve φ.

Braham and Van Hees [7]: This article examines *backwards-responsibility*, predominantly through game theory.

Pipatti [28] Primarily a philosophical paper, this article focusses on group responsibility.

Baldoni, et al [2,3]: These papers mainly tackle organisations and accountability, bringing in aspects such as responsibility *distribution*:

> *"We denote by **A** a set of accountabilities, calling it an accountability specification, and by **R** a responsibility distribution, that is a set of responsibility assumptions that complement the specification of an agent organization."*

Oddie and Tichý [26]: Based on the λ-calculus this complex approach appears to involve assessment of what the agent is able to do and so what it is/was responsible for, in an accountability vein. For example

> *"An agent is (partially) responsible for a state of affairs θ if θ is now inevitable and there was a time at which something the agent could have done would have averted it."*

Collenette, et al [11], Moth-Lund Christensen, et al [25]: Recent articles based on the philosophical work of Whittle [33] and targeting a practical route from the concept of responsibility to practical agent activity. In particular, [11] targets a computational encoding of responsibilities as being 'above' the level of goals and intentions. Responsibilities require a range of diverse activities, such as achievement goals, actions, and runtime verification. Note that detailed logical/axiomatic formalisation is not included in either of these works.

3 Formal Basis for Responsibilities in Agents

We specifically wish to define "responsibility" as a first-class entity in its own right, rather than deriving it from other entities. Furthermore, we wish to see how far a simple formalisation will take us, without resorting to more complex aspects such as *strategic reasoning*, *deontic logics*, *probability*, or *game theory*.

3.1 Informal (and Incomplete) Syntax and Semantics

Let us start with agents and properties:

- Agents: from the finite set $Agent = \{1, 2, \ldots, n\}$
- Atomic property: a standard propositional logic formula, φ

We then assume that the formalisation takes place within a temporal logic context and so will utilise linear temporal logic operators such as $\Box$ ("always in the future"), $\Diamond$ ("sometime in the future"), etc[3] Then, some basic agent operators:

- Belief operator: where $\mathbf{B}_i\varphi$ is true if agent i believes φ to be true
- Intention operator: where $\mathbf{I}_i\varphi$ is true if agent i intends φ to be true
- Knowledge operator: where $\mathbf{K}_i\varphi$ is true if agent i knows φ to be true
- Capability operator: where $\mathbf{C}_i\varphi$ is true if agent i is capable of making φ true

We define two varieties of *responsibility* beginning with "Responsibility for": *an agent being responsible for achieving/maintaining something*

$\mathbf{R}_i\varphi$ is true if agent i is *responsible* for achieving/maintaining φ

Example: "a parent is responsible for ensuring a child is safe".

Then, "Responsibility to": *an agent being responsible to another agent (for some aspect)*

$\mathbf{R}_i^j\varphi$ is true if agent i is *responsible to* agent j (for φ)

Example: "a child is responsible to their parent"[4].

Example: "an employee is responsible to their employer (for their work)"

3.2 Aside: Axioms and Interactions

Once we have a formalisation of one or more (modal) operators then we can examine a whole range of axioms and interactions concerning these operators. We are not going to delve into this aspect in any detail, but below just highlight some obvious axioms.

[3] Note that we will, at times, abuse this notation and also allow branching temporal logic operators such as $\mathbf{A}$ ("for all future paths") and $\mathbf{E}$ ("for some future path")!

[4] If one agent is responsible to another for *everything* we might use $\mathbf{R}_i^j$**true** or just $\mathbf{R}_i^j$

Axioms for *responsibility to*:

- $\vdash \mathbf{R}_i^i$ ..... agent is responsible to itself
- $\vdash (\mathbf{R}_i^j \wedge \mathbf{R}_j^k) \Rightarrow \mathbf{R}_i^k$ .....transitivity
- $\vdash \mathbf{R}_i^j \Rightarrow \neg \mathbf{R}_j^i$ .....mutual responsibility
- $\vdash (\mathbf{R}_i^j \varphi \wedge \mathbf{R}_i^j(\varphi \Rightarrow \psi)) \Rightarrow \mathbf{R}_i^j \psi$
- $\vdash (\mathbf{R}_i^j \varphi \wedge \mathbf{R}_k^i \psi \wedge \Box(\varphi \Rightarrow \psi)) \Rightarrow \mathbf{R}_k^j \psi$

Axioms for *responsibility for*:

- $\vdash \mathbf{R}_i \mathbf{R}_i \varphi \Rightarrow \mathbf{R}_i \varphi$
- $\vdash \mathbf{R}_i \varphi \Rightarrow \mathbf{R}_i \mathbf{R}_i \varphi$ ..... modal axiom '4'?

Negating responsibility: Do we want to insist that, if agent i is responsible for something then it is not responsible for the opposite, i.e:

- $\vdash \mathbf{R}_i \varphi \Rightarrow \neg \mathbf{R}_i \neg \varphi$

or are we happy that an agent can be responsible for two contradictory things, i.e:

- $\vdash \mathbf{R}_i \varphi \wedge \mathbf{R}_i \neg \varphi$ is satisfiable

Aside (options to be explored): If we allow $\mathbf{R}_i \varphi \wedge \mathbf{R}_i \neg \varphi$ then the $\mathbf{R}_i$ operator does not satisfy the **D** modal axiom!

If we do not allow $\mathbf{R}_i \varphi \Rightarrow \mathbf{R}_i \mathbf{R}_i \varphi$ then we also do not have the **4** axiom.

Should we have the modal **5** axiom: $\neg \mathbf{R}_i \varphi \Rightarrow \mathbf{R}_i \neg \mathbf{R}_i \varphi$?

Should we have the modal **B** axiom: $\varphi \Rightarrow \mathbf{R}_i \neg \mathbf{R}_i \neg \varphi$?

We may return to this aspect in the future but, next, we turn to our main aim which is to describe how we move from responsibilities to action/goals in our agents.

4 "Responsibility To"

This operator aims to capture the interdependencies of responsibilities amongst agents. While we are not going to explore this in detail, we note three important elements.

1. We will use the $(\mathbf{R}_i^j$ below only to indicate that one agent is responsible to another. Essentially, we will use this to justify inaction, since another agent (responsible to our agent) should deal with the situation.
2. Axioms such as $(\mathbf{R}_i^j \wedge \mathbf{R}_j^k) \Rightarrow \mathbf{R}_i^k$ help us build up useful structures, not unlike *delegation hierarchies* [15].
3. However, the specialisation to specific properties, for example $(\mathbf{R}_i^j \varphi \wedge \mathbf{R}_j^k \varphi) \Rightarrow \mathbf{R}_i^k \varphi$, makes this *significantly* more complex.

Example: I am responsible for filling in forms, but so is X. As X is responsible to me (at least for this) and as only one person needs to fill in the forms then I don't have to do any form-filling (I believe that X will do this and so I don't have to). But X is on holiday (or X has more important responsibilities) so I now believe I have to fill in the forms.

5 From "Responsibility For" to Activity

We say that our agent will only choose to do (or at least try to do) something if it is *responsible* for achieving that thing, needed for that thing, and capable of achieving that thing. Essentially,

IF agent a is responsible for achieving φ

AND agent a believes at least one more agent is needed for this

AND there is something (ψ) that a is capable of doing that a believes will lead to φ,

THEN agent a adopts a goal to achieve φ.

Before turning to formalisation, let us consider some of these elements.

5.1 Need

As above, a key element is that agent a believes that "at least one more agent is needed for φ". There are two aspects to this:

1. the number of agents committed to achieving φ has not yet reached the number required;
2. our agent might not be needed if sufficient other agents are responsible for φ and our agent is *not* responsible to any of those agents.

Addressing the first aspect

Example: A large box requires 4 robots to lift it. If agent lifter is responsible for ensuring the box is lifted and there are fewer than 4 robots lifting at present, then lifter will try to help lift.

N.B: If 4 robots are already lifting, our robot is not needed[5]!

To simplify this discussion, and the formalisation, we will mostly assume that only one agent is needed in order to achieve φ.

So, now for the second aspect. Even if an agent is responsible for φ then if our agent is *responsible to* that agent, our agent will endeavour to achieve φ. So, if both $\mathbf{R}_i\varphi$ and $\mathbf{R}_j\varphi$ then both, in principle, might try to achieve φ. However, if j is responsible to i, i.e. $\mathbf{R}^i_j$, then only j need do this. Essentially, agent i does not need to tackle φ even though it can.

5.2 Capability

If an agent believes it is capable of achieving φ, and the agent is both responsible and needed, then the agent should aim to achieve φ.

Example: If agent watcher is responsible for ensuring the box is lifted and there are fewer than 4 robots lifting at present, and watcher is capable of undertaking an action, then it does.

N.B: If watcher is not capable of undertaking an action, then it does nothing here.

[5] We also need to refer to the second aspect in case any of those 4 robots anticipate our lifter robot will help as it is responsible to one of the 4.

5.3 Formalisation

So, given the above (and simplifying the capacity constraints so that only one agent is needed to achieve φ) then we formalise the general statement as

IF $\mathbf{R}_a\varphi \wedge$

AND $\forall i.\ (\mathbf{B}_a\mathbf{R}_i\varphi \Rightarrow \mathbf{R}^i_a) \wedge$

AND $\exists \psi.\ \mathbf{C}_a\psi \wedge \mathbf{B}_a\Box(\psi \Rightarrow \Diamond\phi)$

THEN $\mathbf{G}_a\psi$

Clarifications regarding this are

- In $\forall i.\ (\mathbf{B}_a\mathbf{R}_i\varphi \Rightarrow \mathbf{R}^i_a)$ we are formalising that agent a is responsible to all the other agents that are, in turn, responsible for φ. And so a needs to tackle φ.
- Concerning the *capability* clause, a nuance here is that our agent might not be directly capable of achieving φ. However, if there is something (ψ) that the agent is capable of and that it believes will lead to φ, then it should undertake ψ.

5.4 Agents Doing Nothing

The above leads straightforwardly to situations where our agent, a need not do anything:

IF agent a is ***not responsible*** for achieving **THEN** a does not have to do anything

IF a believes other agents ***"have this covered"*** **THEN** a does not have to do anything

IF agent a is ***not capable*** of achieving φ[6]**THEN** a does not have to do anything

6 Lack of Responsibility

Now we will explore what "not being responsible" for something might mean within our context.

6.1 Indifference

What is the negation of "being responsible for" something? An obvious answer is being *irresponsible* but that does not quite work. We could also use *carefree*. However, we have settled on *indifferent*, and use the operator '**N**' for this. So an agent is indifferent to some property if, and only if, it is not responsible for making it true:

$$\mathbf{N}_i\varphi \quad \Leftrightarrow \quad \neg\mathbf{R}_i\varphi$$

We have some further possibilities. For example, the common meaning of "indifference" is that we don't care. And so, if our agent is indifferent to φ then maybe it is neither responsible for φ nor for $\neg\varphi$?

$$\mathbf{N}_i\varphi \quad \Rightarrow \quad \neg\mathbf{R}_i\varphi \wedge \neg\mathbf{R}_i\neg\varphi?$$

[6] Or anything that will lead to φ

Again, once we formalise this in more detail, then indifference should directly affect agent behaviour.

Finally, once we have this indifference operator we can indulge in explorations around various axioms. For example, which of these might make sense??

D: $\mathbf{R}_i\varphi \Rightarrow \mathbf{N}_i\varphi$

4: $\mathbf{R}_i\varphi \Rightarrow \mathbf{R}_i\mathbf{R}_i\varphi$

5: $\mathbf{N}_i\varphi \Rightarrow \mathbf{R}_i\mathbf{N}_i\varphi$

B: $\varphi \Rightarrow \mathbf{R}_i\mathbf{N}_i\neg\varphi$

6.2 From Indifference to Recklessness

We now briefly want to explore this indifference, i.e. lack of responsibility with respect to a particular issue, in combination to with knowledge about the effects of this inactivity. Let us imagine that φ is some significant property we wish to preserve, e.g. "safety of a child". Then we are going to attempt to formally distinguish between "lack of responsibility" concerning this issue, i.e. depending on how much reasoning/knowledge the agent invoked. So, we have three options here

- Agent did not 'think' at all about φ when deciding what to do
  "agent a is **indifferent** *to φ"*
- Agent believed that φ might be violated but continued
  "agent a is **reckless** *with respect to φ"*
- Agent believed that the choice it was making was very likely to lead to φ being violated yet continued
  "agent a is **malicious** *with respect to φ"*

The idea here is to distinguish culpability (and, if the agents were human, liability) around the agent's indifference.

6.3 Formalisation

It is here that we overload our notation and also allow branching temporal logic operators such as $\mathbf{A}$ ("for all future paths") and $\mathbf{E}$ ("for some future path")! We also assume that in all cases our agent is indifferent, in the formal sense above, about φ, i.e. $\mathbf{N}_a\varphi$.

So, we begin with straightforward indifference where our agent decides to do something (by making it an intention) but has not thought about whether this will lead to the violation of our key property (φ). So:

$$\mathbf{I}_a\psi \wedge \neg\mathbf{B}_a(\psi \Rightarrow \mathrm{A}\Box\neg\varphi)$$

Here a didn't believe that the intended action (ψ) would necessarily lead to φ being violated.

Recklessness: here we say that agent a is "reckless" with respect to φ" if it believed φ might be violated but continued regardless:

$$\mathbf{I}_a\psi \wedge \mathbf{B}_a(\psi \Rightarrow \mathrm{E}\Diamond\neg\varphi)$$

Maliciousness: finally, we say agent a is "malicious" with respect to φ" if it believed that the choice it was making was certain[7] to lead to φ being violated

$$\mathbf{I}_a \psi \wedge \mathbf{B}_a(\psi \Rightarrow \mathrm{A}\Diamond\neg\varphi)$$

From a functional point of view φ might be violated in any of the above situations. However, from the point of view of human confidence and trust, the distinction between these options is very significant.

7 Conclusions

This initial work has examined formalisations of responsibility, of various forms, and their impact on agent activity. A motivating factor has been to represent situations in which our agent (or robot) could do something and has some responsibility for doing it, yet legitimately decides to do nothing. We have also looked at different views of "lack of responsibility" and how the agent's beliefs might interact with these to provide a formal representation of concepts such as "recklessness". Throughout our exploration, a core aim has been to retain the smallest (reasonable) logical formulation that appears to make sense.

7.1 Future Work: Formalisation

This is just an initial study and there is much work remaining. Improving the route from logical representation to to agent actions/goals/intentions is clearly a priority. Further exploration of indifference, recklessness, and malice will also be interesting, especially the *positive* versions of these definitions and how these impact upon agent reasoning. For example, if our agent is ***not*** reckless it should carry out some hypothetical reasoning to be clearer about the potential impacts of its actions. Further analysis of the relationship between $\mathbf{R}_b^a$ and these concepts might be needed.

7.2 Future Work: Implementation

Effective implementation, involving further exploration, in computational agent systems is clearly important. As well as embedding suitable concepts of "responsibility" into our practical agent computation we also aim to explore the use of this formalisation in assessing other systems developed in different ways. A promising route appears to be to use the approach in [12] for matching outcomes against a model of unbiased, or fair [13], behaviour. We aim to explore something similar with a model of responsibility (or irresponsibility) as defined in this article, ideally enabling us to detect *irresponsible* behaviour.

7.3 Future Work: Context

The question of where responsibilities come from and how they change over the lifetime of an agent is also important. Though old, we aim to look at group structuring using contexts to supply the dynamic elements and to enforce more complex dynamic organisations. Since, in

[7] Here we might envisage a probabilistic element to modify this to "very likely".

that work, groups of agents are themselves agents, then the "responsible to" and "responsible for" concepts will also be extended to groups of agents.

Acknowledgments: Thanks to Helen Beebee, Joe Collenette, Louise Dennis, Sarah Moth-Lund Christensen, and Ann Whittle for comments on earlier versions of these ideas.

References

[1] R. Alur, T.A. Henzinger & O. Kupferman (2002): *Alternating-Time Temporal Logic*. *Journal of the ACM* 49, pp. 672–713.

[2] M. Baldoni, C. Baroglio, O. Boissier, K.M. May, R. Micalizio & S. Tedeschi (2018): *Accountability and Responsibility in Agent Organizations*. In: *Proc. 21st International Conference on Principles and Practice of Multi-Agent Systems*, *Lecture Notes in Computer Science* 11224, Springer, pp. 261–278, doi:10.1007/978-3-030-03098-8_16.

[3] M. Baldoni, C. Baroglio & R. Micalizio (2019): *Accountability, Responsibility and Robustness in Agent Organizations*. In: *Proc. 34th Italian Conference on Computational Logic*, *CEUR Workshop Proceedings* 2396, CEUR-WS.org.

[4] M. Baldoni, G. Boella & L.W.N. van der Torre (2009): *The Interplay between Relationships, Roles and Objects*. In: *Proc. Third IPM International Conference on Fundamentals of Software Engineering*, *Lecture Notes in Computer Science* 5961, Springer, pp. 402–415, doi:10.1007/978-3-642-11623-0_24.

[5] N. Belnap (1991): *Backwards and Forwards in the Modal Logic of Agency*. *Philosophy and Phenomenological Research* 51(4), pp. 777–807.

[6] N. Belnap & M. Perloff (1992): *The Way of the Agent*. *Studia Logica* 51, pp. 463–484.

[7] M. Braham & M. van Hees (2012): *An Anatomy of Moral Responsibility*. *Mind* 121(483), pp. 601–634, doi:10.1093/mind/fzs081.

[8] M.E. Bratman (1987): *Intentions, Plans, and Practical Reason*. Harvard University Press.

[9] G. Cabri, L. Ferrari & L. Leonardi (2004): *Agent Role-based Collaboration and Coordination: A Survey about Existing Approaches*. In: *Proc. IEEE International Conference on Systems, Man & Cybernetics*, IEEE, pp. 5473–5478, doi:10.1109/ICSMC.2004.1401064.

[10] P.R. Cohen & H.J. Levesque (1990): *Intention is Choice with Commitment*. *Artificial Intelligence* 42, pp. 213–261.

[11] J. Collenette, L.A. Dennis & M. Fisher (2023): *Prospective Responsibility for Multi-agent Systems*. In: *Proc. 43rd SGAI International Conference on Artificial Intelligence*, Springer-Verlag, p. 247–252, doi:10.1007/978-3-031-47994-6_23.

[12] G. Coraglia, F.A. D'Asaro, F.A. Genco, D. Giannuzzi, D. Posillipo, G. Primiero & C. Quaggio (2023): *BRIOxAlkemy: a Bias Detecting Tool*. In: *Proc. 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence*, *CEUR Workshop Proceedings* 3615, CEUR-WS.org, pp. 44–60. Available at `https://ceur-ws.org/Vol-3615/paper4.pdf`.

[13] G. Coraglia, F.A. Genco, P. Piantadosi, E. Bagli, P. Giuffrida, D. Posillipo & G. Primiero (2024): *Evaluating AI fairness in credit scoring with the BRIO tool*. doi:10.48550/ARXIV.2406.03292. arXiv:2406.03292.

[14] L.A. Dennis & M. Fisher (2023): *Verifiable Autonomous Systems: Using Rational Agents to Provide Assurance about Decisions Made by Machines*. Cambridge University Press, doi:10.1017/9781108755023.

[15] M. Dobrajska, S. Billinger & S. Karim (2015): *Delegation Within Hierarchies: How Information Processing and Knowledge Characteristics Influence the Allocation of Formal and Real Decision Authority*. *Organizational Science* 26(3), pp. 687–704, doi:10.1287/ORSC.2014.0954.

[16] M. Fisher, V. Mascardi, K.Y. Rozier, B. Schlingloff, M. Winikoff & N. Yorke-Smith (2021): *Towards a Framework for Certification of Reliable Autonomous Systems*. *Autonomous Agents and MultiAgent Systems* 35(1), p. 8, doi:10.1007/s10458-020-09487-2.

[17] A. Giordani (2018): *Ability and Responsibility in General Action Logic*. In J. Broersen, C. Condoravdi, N. Shyam & G. Pigozzi, editors: *Deontic Logic and Normative Systems*, College Publications.

[18] D. Glavaničová & M. Pascucci (2019): *Formal Analysis of Responsibility Attribution in a Multimodal Framework*. In: *Proc. 22nd International Conference on Principles and Practice of Multi-Agent Systems (PRIMA)*, Springer, pp. 36–51.

[19] V. Koeman, L.A. Dennis, M. Webster, M. Fisher & K. Hindriks (2019): *The"Why did you do that?" Button: Answering Why-questions for end users of Robotic Systems*. In: *Proc. of the 7th International Workshop on Engineering Multi-Agent Systems (EMAS)*, doi:10.1007/978-3-030-51417-4_8.

[20] K. Konolige (1985): *A Computational Theory of Belief Introspection*. In: *Proc. International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, pp. 502–508.

[21] T. de Lima, L.M.M. Royakkers & F. Dignum (2010): *A Logic for Reasoning about Responsibility*. *Logic Journal of the IGPL* 18(1), pp. 99–117, doi:10.1093/jigpal/jzp073.

[22] E. Lorini, D. Longin & E. Mayor (2013): *A Logical Analysis of Responsibility Attribution: Emotions, Individuals and Collectives*. *Journal of Logic and Computation* 24(6), pp. 1313–1339, doi:10.1093/logcom/ext072.

[23] E. Lorini & G. Sartor (2015): *Influence and Responsibility: A Logical Analysis*. In: *Legal Knowledge and Information Systems*, Frontiers in Artificial Intelligence and Applications, IOS Press, pp. 51–60.

[24] R.C. Moore (1995): *Logic and Representation (CSLI Lecture Notes Number 39)*. Center for the Study of Logic and Information (CSLI), Stanford University, USA. (Distributed by Chicago University Press).

[25] S. Moth-Lund Christensen, H. Beebee, L. Dennis & M. Fisher (2024): *Prospective Responsibilities as the Foundation for Agent Decision-Making*. In preparation.

[26] G. Oddie & P. Tichý (1982): *The Logic of Ability, Freedom and Responsibility*. *Studia Logica* 41, p. 227–248, doi:10.1007/BF00370346.

[27] J. Odell, H.V.D. Parunak & M. Fleischer (2002): *The Role of Roles in Designing Effective Agent Organizations*. In: *Software Engineering for Large-Scale Multi-Agent Systems, Research Issues and Practical Applications [the book is a result of SELMAS 2002]*, *Lecture Notes in Computer Science* 2603, Springer, pp. 27–38, doi:10.1007/3-540-35828-5_2.

[28] O. Pipatti (2019): *The Anatomy of Moral Responsibility*. In: *Morality Made Visible*, Routledge.

[29] A.S. Rao & M. Georgeff (1995): *BDI Agents: From Theory to Practice*. In: *Proc. 1st Int. Conf. Multi-Agent Systems (ICMAS)*, San Francisco, USA, pp. 312–319.

[30] A.S. Rao & M.P. Georgeff (1991): *Asymmetry Thesis and Side-Effect Problems in Linear Time and Branching Time Intention Logics*. In: *Proc. IJCAI*, pp. 498–504.

[31] A.S. Rao & M.P. Georgeff (1991): *Modeling Agents within a BDI-Architecture*. In: *Proc. 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR&R)*, Morgan Kaufmann, pp. 473–484.

[32] L. Royakkers & J. Hughes (2020): *Blame it on me*. *Journal of Philosophical Logic* 49, p. 315–349, doi:10.1007/s10992-019-09519-7.

[33] A. Whittle (2024): *An Analysis of Prospective Responsibilities*. *The Journal of Ethics*.

[34] A.F.T. Winfield, S. Booth, L.A. Dennis, T. Egawa, H. Hastie, N. Jacobs, R.I. Muttram, J.I. Olszewska, F. Rajabiyazdi, A. Theodorou, M.A. Underwood, R.H. Wortham & E. Watson (2021): *IEEE P7001: A Proposed Standard on Transparency*. *Frontiers in Robotics and AI* 8, doi:10.3389/frobt.2021.665729.

[35] M. Wooldridge (2002): *An introduction to MultiAgent Systems*. John Wiley and Sons, LTD.

[36] V. Yazdanpanah, M. Dastani, W. Jamroga, N. Alechina & B.S. Logan (2019): *Strategic Responsibility Under Imperfect Information*. In: *Proc. AAMAS*.