Accountability for Practical Reasoning Agents

Stephen Cranefield[1][0000−0001−5638−1648], Nir Oren[2][0000−0002−4854−9014], and
Wamberto Vasconcelos[2][0000−0001−5090−7581]

stephen.cranefield@otago.ac.nz,
n.oren@abdn.ac.uk, w.w.vasconcelos@abdn.ac.uk

[1] University of Otago, Dunedin, New Zealand
[2] University of Aberdeen, Aberdeen, UK

Abstract. Artificial intelligence has been increasing the autonomy of man-made artefacts such as software agents, self-driving vehicles and military drones. This increase in autonomy together with the ubiquity and impact of such artefacts in our daily lives have raised many concerns in society. Initiatives such as transparent and ethical AI aim to allay fears of a "free for all" future where amoral technology (or technology amorally designed) will replace humans with terrible consequences. We discuss the notion of accountable autonomy, and explore this concept within the context of practical reasoning agents. We survey literature from distinct fields such as management, healthcare, policy-making, and others, and differentiate and relate concepts connected to accountability. We present a list of justified requirements for accountable software agents and discuss research questions stemming from these requirements. We also propose a preliminary formalisation of one core aspect of accountability: responsibility.

1 Introduction

Accountability has become an increasingly common term in public discourse, with frequent demands for organisations and officials such as politicians, business leaders, government agencies and public service organisations to be held accountable for their actions (or lack of action). Dubnick [1] describes the term "accountability" as a *cultural keyword*—one that was "culturally innocuous" until the 1960s–70s, but has since undergone a massive growth in usage and become an "expansive, ambiguous, and often enigmatic term with considerable cultural gravitas".

With the increasing capabilities and uptake of machine learning and other AI techniques to aid human decision-making, the public desire for accountability has begun to encompass the development and deployment of AI software [2, 3], and is likely to provide increasing urgency for researchers to address the emerging field of the ethical use of AI [4–6] (see also DeepMind's "Ethics and Society" initiative[3]). Due to the conspicuous success of deep learning classifiers and reinforcement learning systems (e.g., Alphabet's AlphaGo[4]), one particular research focus is on understanding and addressing the inherent biases due to the dependency of such systems on large sets of training data [7]. This is an example of accountability applied to the *people and organisations*

---

[3] https://deepmind.com/applied/deepmind-ethics-society/

[4] https://deepmind.com/research/alphago/

involved in developing and deploying AI: academic (and increasingly public) debate is driving the development and application of norms of best practice [7].

However, in the context of AI systems that can *act autonomously*, the question arises of whether, and how, such systems could themselves be considered as "accountable". This is particularly important for systems that are adaptive, i.e., those that have the flexibility to modify their behaviour-generating processes due to changes in their current knowledge of the world and their interactions with other "agents", which might be humans or other autonomous software systems. This paper addresses the accountability of adaptive autonomous systems, with a particular focus on agents that reason using goals and plans, such as belief-desire-intention (BDI) agents [8–10], which have a long history of investigation by researchers in the field of multi-agent systems.

The contributions of this article are: i) a survey of the relevant literature on accountability, drawing from diverse areas such as sociology, healthcare, management, policy-making and artificial intelligence (especially autonomous and multi-agent systems); ii) a differentiation and correlation among concepts closely connected to accountability such as responsibility, answerability, and others; we also discuss the functional purpose of accountability; iii) a justified list of requirements for accountable autonomous agents and research questions stemming from these; and iv) a preliminary formalisation of one core aspect of accountability: answerability.

The rest of this paper is organised as follows. Section 2 surveys contributions from disparate areas, to answer the question "what is accountability?". Section 3 proposes, based on the literature surveyed, requirements to support accountability in autonomous practical reasoning agents; for each requirement we list associated research questions. In Section 4 we present a preliminary formal model of one aspect of accountability: answerability. We conclude the paper in Section 5, discussing our approach, contributions and further research.

2   What is accountability?

There has been a small amount of prior work related to accountability of autonomous systems, but it is not clear that this work has formed a consensus on what accountability entails, or how well that work aligns with the view of accountability in other academic fields. Therefore, in this section we survey the literature on accountability from disparate fields such as policy-making, sociology, management and computing science (especially artificial intelligence and multi-agent systems). Our aim is to identify the key requirements that an autonomous agent would need to satisfy in order to be considered accountable.

Chopra and Singh [11] describe accountability as a normative concept in the context of socio-technical systems: "accountability requirements describe how principals ought to act in each other's eyes, providing a basis for their mutual expectations". They give two examples of accountability requirements: a meeting participant who is accountable for turning up to a meeting after accepting an invitation, and a food company that is accountable to a regulator for maintaining certain tracking information and providing it to a regulator on demand. However, it is not clear from this discussion to what degree (if any) the authors believe the computational representations and processes needed to

support accountability might differ from existing techniques developed by multi-agent systems researchers for reasoning about norms and commitments [12, 13].

Baldoni et al. [14] propose the study of *computational accountability*. They consider accountability to be an ethical value, and define accountability as "the acknowledgment and assumption of responsibility for decisions and actions that an individual, or an organization, has towards another party". They note that, implicitly, "individuals are expected to account for their actions and decisions when put under examination". The paper focuses on multi-agent systems that track the state of conditional social commitments using business artifacts, in order to "coordinate their activities, e.g. through responsibility assignment, as well as to identify liabilities". It is argued that the "analysis of accountability can be accomplished by looking at commitment relationships".

In later work, Baldoni et al. [15, 16] take the viewpoint of *accountability as a mechanism*, summarised by Bovens et al. [17] as "an institutional relation or arrangement in which an agent can be held to account by another agent or institution". They consider how such an institutional mechanism can be provided by design in a multi-agent system (MAS), and seek to provide "structures that allow assessing who is accountable without actually infringing on the individual and private nature of agents" and to "determine action impact or significance by identifying the amount of disruption it causes in terms of other agents and/or work affected" [15]. To this end, they present five "necessary-but-not-sufficient principles that an MAS system must exhibit in order to support accountability determination" [15]. These principles state that (i) agents should interact within the scope of an organisation, (ii) must join the organisation by taking on a role, (iii) can be accountable only for goals they have explicitly accepted, and (iv) may specify the resources they need to satisfy a goal (which may be provided, or not, at the organisation's discretion). The fourth principle is endowed with particular significance for accountability determination: "Should an uniformed agent stipulate insufficient provisions for an impossible goal that is then accepted by an organization, that agent will be held accountable because by voicing its provisions, it declared an impossible goal possible" [16]. Baldoni et al. operationalise these principles as an "accountability protocol" to be followed when an agent joins an organisation. This protocol ensures the creation of specific types of commitment between agents and between agents and the organisation. This work is situated within a particular paradigm of organisational multi-agent systems in which organisations are supported by specialised coordination artifacts, whereas we seek a more general model of computational accountability.

Dignum [18] addresses the question of how AI systems can be designed responsibly to ensure they are "sensitive to moral principles and human value [sic]". She discusses three principles of responsible AI: accountability, responsibility and transparency (ART). Accountability is described as "the need to explain and justify one's decisions and actions to its partners, users and others with whom the system interacts". In addition, there is a need for moral values and social norms to be represented and included in the system's deliberations and explanations of its decisions.

Other multi-agent systems researchers have investigated related concepts such as responsibility, which we discuss in Section 2.1, after a more general look at the literature on accountability.

Dubnick [1] notes that it is difficult to find a definition of accountability that is not circular or specific to a qualifying adjective (e.g. "political accountability"). In the latter case, Dubnick observes that "whatever substantive meaning might be in the word accountability is overwhelmed and subordinated to the demands of the specific task environment". Fox [19] also notes the lack of clarity around the meaning of accountability and related concepts, stating that "the terms *transparency* and *accountability* are both quite malleable and therefore – conveniently – can mean all things to all people".

Bovens et al. [17] discuss the views of accountability in the social psychological, accounting, public administration, political science, international relations and constitutional law literature. They observe that there is a "minimal consensus" in the academic literature. Schillemans [20] expresses this consensus as follows:

1) Accountability is about providing answers, about answerability, towards others with a legitimate claim in some agents' work. 2) Accountability is furthermore a relational concept: it focuses our attention on agents who perform tasks for others . . . . 3) Accountability is retrospective . . . and focuses on the behavior of some agent in general, ranging from performance and results to financial management, regularity or normative and professional standards. 4) . . . accountability consists of three analytically distinct phases. In the first phase, the agent/accountor/actor renders an account on his conduct and performance to a significant other. This may be coined the information phase. In the second phase, the principal/accountee/forum assesses the . . . transmitted information and both parties often engage in a debate on this account. The principal/accountee/forum may ask for additional information and pass judgment on the behaviour of the agent/accountor/actor. The agent/accountor/actor will then answer to questions and if necessary justify and defend his course of action. This is the debating phase. Finally, the principal/accountee/forum comes to a concluding judgment and decides whether and how to make use of available sanctions. This is the sanctions or judgment phase.

From this, we note that accountability revolves around some form of *accountability relationship* between an accountee and accountor. As discussed in Section 3.1, many of the properties of this relationship have not yet been formalised.

Emanuel and Emanuel [21] give a definition of accountability in the domain of healthcare: "Accountability . . . entails procedures and processes by which one party provides a justification and is held responsible for its actions by another party that has an interest in the actions". They consider the following components of accountability: the locus of accountability, i.e. *who* can be held accountable, the *domain* of accountability, i.e. for what activities, practices or issues "a party can legitimately be held responsible and called on to justify or change its action", and the *procedures* of accountability, divided into evaluation of compliance and dissemination of evaluations to seek "responses or justifications" from accountable parties.

2.1 Related concepts

Dubnick [1, Fig. 2.4] categorises various concepts related to accountability that are motivated by "moral pull" (i.e., due to external forces): liability, answerability, respon-

sibility, responsiveness (in the legal, organisational, professional and political settings, respectively), and those motivated by "moral push" (i.e., due to internal managerial efforts): obligation, obediance, fidelity, amenability (in the same four settings, respectively).

The relationships between accountability, responsibility and answerability seem especially subject to varying viewpoints. Dubnick [1] notes that one can be "responsible for some event, for example the marriage of two people who met because (one) did not take the empty seat between them on the bus, without being held to account for it". Eshleman [22] discusses various philosophical views on *moral* responsibility. The *accountability* view holds that "an agent is responsible, if and only if it is appropriate for us to hold her responsible, or accountable, via the reactive attitudes ... (e.g. resentment)". Another influential view, referred to by Eshleman as the *answerability* view, is that "someone is responsible for an action or attitude just in case it is connected to her capacity for evaluative judgment in a way that opens her up, in principle, to demands for justification from others".

In the practice of business management, a Responsible, Accountable, Consulted, and Informed (RACI) matrix is a recognised [23] tool to map where responsibility and accountability are assigned for activities. In this context, the responsible parties are those who work on the activity (responsibility may be shared), whereas the accountable party is the (unique) person with "yes or no authority" over the activity and "about whom it is said 'The buck stops here' " [24].

Researchers in multi-agent systems and deontic logic have addressed the concept of responsibility as the problem of assigning blame for failures of group plans or norms [25–36]. This problem has been well studied in the literature, and as determining responsibility is a process performed by a principal, it is largely orthogonal to our focus in this paper: the capabilities needed for an accountable agent to play its role in an accountability relationship with a principal. Therefore, we do not attempt to summarise the literature on responsibility as blame assignment.

In the context of the responsible development of AI systems, Dignum [18] defines *transparency* as "the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and to the governance of the data used or created". Fox [19] discusses the relationship between transparency and accountability in human institutions, which is conventionally expressed as "transparency generates accountability". After reviewing the empirical literature he concludes that transparency is necessary for accountability, but far from sufficient. In particular, his analysis shows that "opaque transparency" (limited to providing access to information) does not necessarily result in accountability, whereas an overlap between transparency and accountability occurs when there is answerability, i.e. the capacity or right to demand answers. However, answerability without consequences (e.g. sanctions) is a "soft" form of accountability. To guarantee "hard accountability" (answerability plus consequence, such as sanctions), the intervention of other "public sector actors" is needed.

Winikoff [37] considers the question of the *trustability* of autonomous systems, i.e., how humans can come to trust them, and proposes three prerequisites for such trust: there should be a social framework for recourse; if the system makes a decision with

negative consequences for the user, the system should be able to explain its behaviour; and the system should be subject to verification and validation to give assurance that key behavioural properties hold.

2.2 The functional purpose of accountability

When setting out to design accountable software agents it is important to consider the functional purpose of accountability. Is accountability simply something that satisfies a human desire to feel empowered (even if there is no other effect), or are there some system-level benefits? In the former case, there may be no point in creating accountable agents unless they are interacting with people or other agents. In the latter case, it is necessary to identify the benefits that we wish our agents (or their society) to enjoy.

The purpose of accountability has been analysed in the human context. Bovens provides this commentary [38]:

> "So why is accountability important? . . . In the academic literature and in policy publications about public accountability, three answers recur, albeit implicitly, time and again. Accountability is important to provide a democratic means to monitor and control government conduct, for preventing the development of concentrations of power, and to enhance the learning capacity and effectiveness of public administration."

The first and last of these answers seem most relevant to software agents (assuming that our agents are not power-seeking). The first reason (control) is also noted by Mulgan [39]:

> "The core sense of accountability is clearly grounded in the general purpose of making agents or sub-ordinates act in accordance with the wishes of their superiors. Subordinates are called to account and, if necessary, penalized as means of bringing them under control."

We note that this also highlights a *motivational* aspect of accountability: a rational agent (as software agents are generally designed to be) will be likely to prioritise goals for which it is accountable, and devote more resources to them. This is due to the expected costs of requests for answers and possible sanctions in the event of sub-standard performance or failure.

Bovens elaborates on the third reason above (enhancing learning) as follows:

> "The purpose of public accountability is to induce the executive branch to learn. The possibility of sanctions from clients and other stakeholders in their environment in the event of errors and shortcomings motivates them to search for more intelligent ways of organising their business. Moreover, the public nature of the accountability process teaches others in similar positions what is expected of them, what works and what does not."

The last sentence implies a norm-alignment and spreading function of accountability, as Bovens notes elsewhere in his article: "Norms are (re)produced, internalised and, where necessary, adjusted through accountability".

We conclude that for (software) multi-agent systems, accountability has a role to play in motivating good performance, and in monitoring and control (when one agent is a subordinate of another). It can also allow for incremental system improvement through learning or instruction, e.g. one agent may send new plans to another agent as an outcome of an accountability dialogue, and can enable the alignment and spreading of norms. When human users or partners are involved, we also see accountability contributing to the alignment of values.

3 Requirements for accountable autonomous agents

Based on the literature discussed above, we propose that in order to support accountability, an autonomous practical reasoning agent should have the following four properties:

Expectation-aware. The agent should be able to understand when it becomes subject to the expectations of others, for example through norms and commitments, such as the obligation to provide answers to accountability queries. It should also expect to be held to account, and possibly incur a sanction, after poor performance and failure—this provides the motivation to perform well. Its practical reasoning should be informed by these expectations. This property is likely to be crucial in ensuring that the following two properties are exercised correctly.

Answerable. The agent should be able to answer retrospective queries about its decision-making, within some pre-established scope. These queries may not be made immediately, so it must maintain sufficient information about its past reasoning to enable these queries to be answered. Note that answerability is similar to the concept of explainability, but includes the relational aspects of accountability: an accountable agent is answerable to a specific party that may send queries within some (possibly limited) scope, and these must be answered.

Argumentative. Full accountability cannot be achieved by one-off queries alone. To enable accountability processes to lead to system improvement (including norm and value alignment), an accountable agent should be capable of undertaking extended accountability dialogues in which beliefs, plans, norms and values are challenged, justified and further queried.

Meta-cognitive. The agent must be able to adapt its reasoning mechanisms as a result of accountability dialogues. For example, an agent may need to update its plans, its plan selection mechanism, its failure-handing mechanism, its norms, or its values as a result of advice from its principal. The ability of an agent to alter its own decision-making components is known as meta-cognition [40], although we do not require the agent to monitor its own cognition, but rather to make changes when required by accountability mechanisms.

Additionally, when the scope of accountability includes actions that affect people, the following property is also required:

Value-aware. The agent should maintain information about the relative importance of human values to its organisation or human partner(s) or client(s), and take these into account during its reasoning [41]. This in line with Dignum's ART model of responsible AI [18].

3.1 Research questions

Various research questions stem from the requirements above. When extending autonomous agents to meet the requirements, we have:

Expectation-aware. Research on norm-aware planning in BDI agents, e.g., [42], indicate that it is desirable and possible to extend a standard practical reasoning mechanism to address normative concerns. Our research questions are
  – What practical reasoning approach is most appropriate to be extended with expectations stemming from accountability relationships?
  – What is the minimal information required to enable expectation-aware behaviour in autonomous agents?
  – What game-theoretic aspects are there when agreeing (or not) to be accountable for something?

Answerable. There is a wealth of research on summarising and presenting data and information to different stakeholders, e.g., [43, 44]. We anticipate queries to refer to rich and comprehensive records of decision-making processes and their rationale. Some questions arising are
  – What knowledge/information should be represented to support accountability?
  – What extensions/adaptations are required in the decision-making process(es) to ensure the knowledge and information of the previous question is adequately represented?
  – What kinds of queries should be supported in accountability relations?

Argumentative. Research on formal argumentation has matured and has been applied to many contexts/domains [45]. Some issues arising are:
  – Which formal argumentation techniques can be (re-)used, adapted or extended in the context of accountability queries and how this can be done?
  – How can accountable behaviour (stemming from practical reasoning) be combined/extended with argumentation capabilities?
  – How can argumentation interactions support and affect accountable behaviour (stemming from practical reasoning)?

Meta-cognitive. Multi-agent plan selection and revision have been explored through different approaches (e.g., [41, 46, 47]) indicating that practical reasoning must tackle meta-cognitive issues – agents not only build and follow plans, but they must also reconsider/revisit decisions and reason about the actual decision processes. Some questions arising are:
  – Is there a need for many levels of meta-cognition, whereby agents become aware about being aware about being aware and so on, or would a single meta-cognition level suffice?
  – Would meta-interpretation [48,49] be an adequate and flexible approach to both meta-cognition and answerability?
  – Should practical reasoning always embed meta-cognitive concerns or should these only be addressed when agents are accountable for some behaviour or result?

Value-aware. Accountable agents seek to act, or answer queries, in a manner which promotes the values of the organisation(s), human partner(s) or client(s) to which they are accountable. In this context, research questions include

- How can the actions for which one is held accountable be shown to align to the values that should be promoted? Existing work on argument based practical reasoning (e.g., [50]) demonstrates the links between action and values, but not between accountability and values.
- How can the lack of promotion of a value (e.g., due to the sub-standard execution of a task) trigger the accountability process?

4 Towards a formalisation of accountability

In this section we propose an initial high-level formalism of accountability, focusing on answerability. We assume the accountable agent is equipped with a well studied form of expectation-awareness: the ability to represent and perform practical reasoning informed by norms such as obligations [51]. We consider that answerability is naturally expressed as an organisational norm, or as a commitment (if implicitly created via a *commitment protocol* [52]). We focus here on the normative view and model answerability as a conditional obligation norm. It is not the intention of this paper to define or commit to any specific formalisation for obligations, so for brevity we use an existing notation from the literature: the logic of Dignum et al. [53] for specifying temporal deontic constraints[5].

$$answerable(ag, at, QL, S, \delta t, rt) \equiv$$
$$\forall q \, PREV(ask(at, ag, q)) \land in\_scope(q, QL, S) \longrightarrow$$
$$O(valid\_reply(ag, at, q, S, \delta t) < now + rt)$$

where:

- ag and at refer to the *account-giver* (or accountable party) and the *account-taker* (or principal), following the terminology of Chopra and Singh [11].
- QL is an agreed (or imposed) query language in which accountability queries will be expressed.
- S is an agreed (or imposed) scope of queries—not all queries that can be expressed in QL may be relevant to the accountability relationship. Restrictions might include the types of goal considered, and the roles under which the queried activities are performed. We make no commitment regarding how S is expressed.
- δt is the length of the retrospective time period that accountability queries can ask about (where $\delta_t = \infty$ means there is no limit). This limits the time interval for which ag must keep records of its decision-making processes.
- rt is the maximum time allowed for an answer to an accountability query to be sent.
- $PREV(a)$ means that the action leading to the current state was a.
- $ask(at, ag, q)$ is the action of at asking ag the query q.
- $in\_scope(q, QL, S)$ denotes the condition that the query q is expressed in the query language QL and is within the scope S.

---

[5] This formalism is based on dynamic logic, but it is out of scope of this paper to describe the semantics. Also, note that our purpose here is to *specify* the nature of the obligation implied by answerability. For implementing accountability processes, it is likely that agents can use less expresssive and possibly more specialised, representations of their obligations.

- $O(a < t)$ denotes the obligation for action a to be done before time t.
- $valid\_reply(ag, at, q, S, \delta t)$ is the action of ag sending at a valid answer for query q within scope S, based on a trace of its reasoning for the last δt time units. We do not attempt, within this obligation, to specify the notion of a valid reply. Instead, we consider this an abstract action, and assume that ag and at have a common understanding of what *counts as* [54] a valid reply. Below we propose one option.
- now is a special variable used in the logic of Dignum et al. [53] to refer to the time at which the obligation's conditions become true.

We now consider what could count as a valid answer to the query. An answerable agent should be obliged to provide information about its practical reasoning that led to the queried behaviour, and that is relevant to the query. Before formalising this, we define some notation.

- $\tau_{ag}^{[t-\delta t, t]}$ denotes a full trace of the agent ag's reasoning during the interval $[t-\delta t, t]$. As well as recording successful plan executions, this trace must include information about options considered and not selected, and action and plan failures.
- Given a full trace τ, we write $\tau \upharpoonright q, S$ to denote the restriction of the trace to contain only information relevant to the query q and scope S, and omit S if there is no scope restriction. We leave as an open question whether such a notion of relevance can be defined—if not, $\tau \upharpoonright q, S = \tau$.

We assume that queries are expressed declaratively, with answers returned as variable bindings (or $\bot$ to indicate failure), and that the trace is viewed as a set of facts, and can therefore be decomposed into disjoint sets of facts. We then propose the following conditions for a query reply to be considered valid (where σ ranges over variable substitutions and $\uplus$ denotes disjoint union):

$$\nexists \sigma : (\tau_{ag}^{[t-\delta t, t]} \upharpoonright q, S \models \sigma(q)) \longrightarrow$$
$$reply(ag, at, q, \bot) \text{ counts\_as } valid\_reply(ag, at, q, S, \delta t)$$

$$\tau_{ag}^{[t-\delta t, t]} \upharpoonright q, S \models \sigma(q) \ \wedge$$
$$\tau_{ag}^{[t-\delta t, t]} \upharpoonright q, S = reasons \uplus rest \ \wedge \ rest \not\models \sigma(q) \longrightarrow$$
$$reply(ag, at, q, \langle \sigma, reasons \rangle) \text{ counts as } valid\_reply(ag, at, q, S, \delta t)$$

The first clause states that a reply containing $\bot$ is valid if the query cannot be answered using the time- and scope-restricted trace. The first line of the second clause expresses the condition that the answer is correct, i.e. $\sigma(q)$ is entailed by the scope- and time-restricted trace. The second line first extracts a set of *reasons* from the trace, to help justify the query result, and then requires that at least some of the reasons provided in the answer are *necessary* for the truth of the answer—removing them from the trace would not allow the query to be answered. When these conditions hold, a reply containing the substitution, i.e. a set of variable bindings, and the reasons is considered valid. This notion of a valid answer does not fully specify the reasons that should be given to justify the answer. We believe these will be domain- and context-dependent, and in

general, we envisage the need for a dialogue between the two agents to build up mutual information through a series of queries.

The use of $\tau_{ag}^{[t-\delta t,t]}$ above implies that ag should give an answer that is correct with respect to the *full trace* over the required retrospective time window. However, that does not necessarily mean that ag must actually record the full trace as implied by its semantics. Given a query scope S, it may be possible to answer queries within that scope using a subset of the information in $\tau_{ag}^{[t-\delta t,t]}$. We explain this intuition by using the notion of an *abstraction* of a transition system. We can view the full trace as a transition system on time-stamped agent internal states (but note that the transitions must include the evaluation of failed reasoning rule conditions, as well as successes). Answering queries with a subset of information means reasoning with an abstraction of the transition system [55], which is defined over information states that are (potentially lossy) *projections* of the full agent states.

For a projection function f and a trace τ, we denote the abstracted transition system that f induces by τ^f. The task for the account-giver (or its designer) is then, given a scope S, to find a projection function f_S such that the following property holds:

$$\forall q : in\_scope(q, S), \forall \tau \in Traces, \forall \sigma, (\tau \upharpoonright q,S) \models \sigma(q)) \iff (\tau^{f_S} \upharpoonright q \models \sigma(q))$$

This states that answering queries within scope S by projecting traces using f_S produces the same answers as would be obtained using scope-restricted traces.

This model of answerability opens a number of research directions, including the following:

- There is a need to underpin the notation above with a formal model of agent reasoning. In the context of debugging BDI agent programs, Winikoff [49] provides such a model in the context of debugging agents by asking "why?" and "why not?" questions, which are answered using traces of agent reasoning. His formalism provides much of what is needed here. However, some aspects of this approach may not suit the problem of answerability. For example, queries may be asked some time after the computation in question was run, especially in the case of suboptimal outcomes or failures, and the account-taker may only have partial observability of the agent trace when asking its queries. Also, Winikoff's semantics assume that new beliefs can be semantically associated with the actions they were consequences of. In practice, the world is more complicated: actions can have various degrees of success and failure, and their effects can vary accordingly. Also, the effects may not always be immediately observable. To cater for these complexities, a richer domain model may be needed, and explanations may need to be contingent on the most likely causes of observations.
- A range of useful notions of query language and scope should be investigated. Winikoff investigated questions seeking reasons for why, at a given point of execution, plan steps were or were not performed, or specific conditions were or were not believed. These could be extended to consider extended models of agent reasoning, e.g., those incorporating norms [51] and values [41]. Another potentially useful query type when the account-taker lacks the full trace is "could you have performed X?" for a plan or action X. For argumentative agents, the notion of a

query language should be extended to include assertions such as "P would have been a better plan to choose".

– The problem of choosing a projection function f_S given a scope S is important to ensure that agents only need to record the minimal required information. Also, there is the inverse question of what scope of queries can be answered by an agent that keeps a specific type of audit trail.

5 Conclusions, Discussion and Future Work

This paper surveyed the meaning and purpose of accountability in many areas, connecting and differentiating it from closely related concepts such as responsibility and transparency, among others. We identify the functional purpose of accountability: it enables monitoring and control of self-interested agents of a multi-agent system, and facilitates incremental improvements in the system. The improvement comes about as agents, aware of what they are accountable for, factor this in their choices of autonomous behaviour; the interactions among agents as they query and answer each other (this being guided by their accountability relations) will enable sharing of "best practices" (plans which withstand scrutiny and criticism), whilst aligning and spreading global norms. We have put forward requirements for accountable practical reasoning agents, and for each of these requirements we listed related research questions. We sketched a formalisation for one aspect of accountability: answerability, as part of an investigation into the normative constructs, the information model and reasoning mechanisms necessary for accountable practical reasoning.

Concerns about advances in AI and their impact in society have caught the attention of the media, governments and people in general. AI, coupled with autonomous behaviour, has immense potential, and initiatives have championed ethical and responsible principles for systems and their design. We hope we have made a step towards accountable autonomy, whereby the design and execution of practical reasoning agents is influenced by accountability. Ultimately, this paper aims to increase awareness among the multi-agent systems and software agents community of accountability and related ethical matters in our research. We would also like to consider this paper as a call-to-arms: we can, as a community, and building on the wealth of our research, lead the AI community in this quest for ethical and responsible AI.

In addition to the various research questions raised in previous sections, we are currently extending BDI practical reasoning technologies to explore accountability issues. We are also developing our formalisation of accountability, especially its connections with normative aspects as well as norm-aware BDI reasoning.

References

1. Melvin J. Dubnick. Accountability as a cultural keyword. In Bovens et al. [56].
2. Paul Billingham and Alastair Colin. The democratisation of accountability in the digital age: Promise and pitfalls. Winner of Robert Davies Essay Competition 2016, Skoll Centre for Social Entrepreneurship, Saïd Business School, The University of Oxford, U.K., 2016. Available at https://www.sbs.ox.ac.uk/sites/default/files/Skoll_Centre/Docs/Accountability_BillinghamColin-Jones.pdf.

3. Sandra Wachter. Towards accountable A.I. in Europe? The Alan Turing Institute, U.K., https://www.turing.ac.uk/blog/towards-accountable-ai-europe. Last accessed 2018-07-25.

4. Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. In Keith Frankish and William M.Editors Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence*, page 316–334. Cambridge University Press, 2014.

5. Virginia Dignum. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20(1):1–3, 2018.

6. Tom Simonite. Tech firms move to put ethical guard rails around AI. *Wired*, May 2018. Available at https://www.wired.com/story/tech-firms-move-to-put-ethical-guard-rails-around-ai/ (last accessed 2018-07-29).

7. James Zou and Londa Schiebinger. AI can be sexist and racist — it's time to make it fair. *Nature*, 559:324–326, 2018.

8. Michael P. Georgeff, Barney Pell, Martha E. Pollack, Milind Tambe, and Michael Wooldridge. The Belief-Desire-Intention model of agency. In *Proceedings of the 5th International Workshop on Agent Theories, Architectures, and Languages (ATAL'98)*, pages 1–10, London, UK, UK, 1999. Springer-Verlag.

9. Felipe Rech Meneguzzi, Avelino Francisco Zorzo, and Michael da Costa Móra. Propositional planning in BDI agents. In *Proceedings of the ACM Symposium on Applied Computing*, pages 58–63, New York, NY, USA, 2004. ACM.

10. Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In *Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS'95)*, pages 312–319. AAAI, 1995. https://www.aaai.org/Papers/ICMAS/1995/ICMAS95-042.pdf.

11. A. K. Chopra and M. P. Singh. The thing itself speaks: Accountability as a foundation for requirements in sociotechnical systems. In *2014 IEEE 7th International Workshop on Requirements Engineering and Law*, page 22. IEEE, 2014.

12. Mehdi Dastani, Leendert van der Torre, and Neil Yorke-Smith. Commitments and interaction norms in organisations. *Autonomous Agents and Multi-Agent Systems*, 31(2):207–249, Mar 2017.

13. Nicoletta Fornara and Marco Colombetti. Representation and monitoring of commitments and norms using OWL. *AI Communications*, 23(4):341–356, December 2010.

14. Matteo Baldoni, Cristina Baroglio, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. Computational accountability. In *Proceedings of the AI*IA Workshop on Deep Understanding and Reasoning: A Challenge for Next-generation Intelligent Agents*, volume 1802 of *CEUR Workshop Proceedings*, pages 56–62. CEUR-WS.org, 2017.

15. Matteo Baldoni, Cristina Baroglio, Katherine M. May, Roberto Micalizio, and Stefano Tedeschi. ADOPT JaCaMo: Accountability-driven organization programming technique for JaCaMo. In *PRIMA 2017: Principles and Practice of Multi-Agent Systems*, pages 295–312. Springer, 2017.

16. Matteo Baldoni, Cristina Baroglio, and Roberto Micalizio. The AThOS project: First steps towards computational accountability. In *Proceedings of the 1st Workshop on Computational Accountability and Responsibility in Multiagent Systems*, volume 2051 of *CEUR Workshop Proceedings*, pages 3–19. CEUR-WS.org, 2018.

17. Mark Bovens, Thomas Schillemans, and Robert E. Goodin. Public accountability. In Bovens et al. [56].

18. Virginia Dignum. Responsible artificial intelligence: designing AI for human values. *ITU journal: ICT Discoveries*, 1(1):1–8, 2018.

19. Jonathan Fox. Development in practice. *The uncertain relationship between transparency and accountability*, 17(4-5):663–671, 2007.

20. Thomas Schillemans. The public accountability review: A meta-analysis of public accountability research in six academic disciplines. Working paper, Utrecht University School of Governance, https://dspace.library.uu.nl/handle/1874/275784, 2013.

21. E. J. Emanuel and L. L. Emanuel. What is accountability in health care? *Annals of Internal Medicine*, 124(2):229–239, 1996.

22. Andrew Eshleman. Moral responsibility. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

23. PMI. *Guide to the Project Management Body of Knowledge (PMBOK®Guide)*. Project Management Institute, 5th, 2013.

24. J. Mike Jacka and Paulette J. Keller. *Business Process Mapping: Improving Customer Satisfaction*. Wiley, 2nd edition, 2009.

25. Davide Grossi, Frank Dignum, Lambèr M. M. Royakkers, and John-Jules Ch. Meyer. Collective obligations and agents: Who gets the blame? In Alessio Lomuscio and Donald Nute, editors, *Deontic Logic in Computer Science*, pages 129–145. Springer, 2004.

26. Roberto Micalizio, Pietro Torasso, and Gianluca Torta. On-line monitoring and diagnosis of multi-agent systems: A model based approach. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 848–852. IOS Press, 2004.

27. Cees Witteveen, Nico Roos, Roman van der Krogt, and Mathijs de Weerdt. Diagnosis of single and multi-agent plans. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 805–812. ACM, 2005.

28. Davide Grossi, Lambèr Royakkers, and Frank Dignum. Organizational structure and responsibility. *Artificial Intelligence and Law*, 15(3):223–249, 2007.

29. Femke de Jonge, Nico Roos, and Cees Witteveen. Primary and secondary diagnosis of multi-agent plan execution. *Autonomous Agents and Multi-Agent Systems*, 18(2):267–294, 2009.

30. Rosja Mastop. Characterising responsibility in organisational structures: The problem of many hands. In *Deontic Logic in Computer Science, 10th International Conference, DEON 2010, Fiesole, Italy, July 7-9, 2010. Proceedings*, volume 6181 of *Lecture Notes in Computer Science*, pages 274–287. Springer, 2010.

31. Tiago De Lima, Lambèr M. M. Royakkers, and Frank Dignum. Modeling the problem of many hands in organisations. In *Proceedings of the 19th European Conference on Artificial Intelligence*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 79–84. IOS Press, 2010.

32. Nils Bulling and Mehdi Dastani. Coalitional responsibility in strategic settings. In João Leite, Tran Cao Son, Paolo Torroni, Leon van der Torre, and Stefan Woltran, editors, *Computational Logic in Multi-Agent Systems*, pages 172–189. Springer, 2013.

33. Roberto Micalizio and Pietro Torasso. Cooperative monitoring to diagnose multiagent plans. *Journal of Artificial Intelligence Research*, 51:1–70, 2014.

34. E. Lorini, D. Longin, and E. Mayor. A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6):1313–1339, 2014.

35. Huib Aldewereld, Virginia Dignum, and Wamberto W. Vasconcelos. Group norms for multi-agent organisations. *ACM Transactions on Autonomous and Adaptive Systems*, 11(2):15:1–15:31, 2016.

36. Natasha Alechina, Joseph Y. Halpern, and Brian Logan. Causality, responsibility and blame in team plans. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, pages 1091–1099. IFAAMAS, 2017.

37. Michael Winikoff. Towards trusting autonomous systems. In *Engineering Multi-Agent Systems*, number 10738 in Lecture Notes in Computer Science, pages 3–20. Springer, 2018.

38. Mark Bovens. Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4):447–468, 2007.

39. Mulgan Richard. 'accountability': An ever-expanding concept? *Public Administration*, 78(3):555–573, 2000.

40. Michael L. Anderson and Donald R. Perlis. Logic, self-awareness and self-improvement: the metacognitive loop and the problem of brittleness. *Journal of Logic and Computation*, 15(1):21–40, 2005.

41. Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in BDI agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 178–184. ijcai.org, 2017.

42. Felipe Meneguzzi, Odinaldo Rodrigues, Nir Oren, Wamberto W. Vasconcelos, and Michael Luck. BDI reasoning with normative considerations. *Engineering Applications of Artificial Intelligence*, 43:127 – 146, 2015.

43. Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications*, 22(3):153–186, August 2009.

44. Catherine Mulwa, Seamus Lawless, Mary Sharp, and Vincent Wade. The evaluation of adaptive and personalised information retrieval systems: A review. *International Journal of Knowledge and Web Intelligence*, 2(2/3):138–156, December 2011.

45. Floris Bex, Floriana Grasso, Nancy Green, Fabio Paglieri, and Chris Reed. *Argument Technologies: Theory, Analysis, and Applications*. Studies in Logic and Argumentation. College Publications, 2017.

46. Natasha Alechina, Mehdi Dastani, Brian Logan, and John-Jules Ch. Meyer. Reasoning about plan revision in BDI agent programs. *Theoretical Computer Science*, 412(44):6115 – 6134, 2011.

47. J. Ma, W. Liu, J. Hong, L. Godo, and C. Sierra. Plan selection for probabilistic BDI agents. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, pages 83–90, Nov 2014.

48. Michael Winikoff. An AgentSpeak meta-interpreter and its applications. In *Programming Multi-Agent Systems, Third International Workshop, ProMAS 2005*, volume 3862 of *Lecture Notes in Computer Science*, pages 123–138. Springer, 2006.

49. Michael Winikoff. Debugging agent programs with "why?" questions. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, pages 251–259. IFAAMAS, 2017.

50. Katie Atkinson and Trevor J. M. Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10-15):855–874, 2007.

51. Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leendert W. N. van der Torre, editors. *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2013.

52. Ashok U. Mallya and Munindar P. Singh. An algebra for commitment protocols. *Autonomous Agents and Multi-Agent Systems*, 14(2):143–163, 2007.

53. F. Dignum, H. Weigand, and E. Verharen. Meeting the deadline: On the formal specification of temporal deontic constraints. In Zbigniew W. Raś and Maciek Michalewicz, editors, *Foundations of Intelligent Systems*, pages 243–252. Springer, 1996.

54. J.R. Searle. *The Construction of Social Reality*. Free Press, 1995.

55. Alain Finkel, S. Purushothaman Iyer, and Grégoire Sutre. Well-abstracted transition systems: application to FIFO automata. *Information and Computation*, 181(1):1–31, 2003.

56. Mark Bovens, Robert E. Goodin, and Thomas Schillemans, editors. *The Oxford Handbook of Public Accountability*. Oxford University Press, 2014.