ABSTRACT
To develop and effectively deploy Trustworthy Autonomous Systems (TAS), we face various social, technological, legal, and ethical challenges in which different notions of responsibility can play a key role. In this work, we elaborate on these challenges, discuss research gaps, and show how the multidimensional notion of responsibility can play a role to bridge them. We argue that TAS requires operational tools to represent and reason about responsibilities of humans as well as AI agents. We review major challenges to which responsibility reasoning can contribute, highlight open research problems, and argue for the application of multiagent responsibility models in a variety of TAS domains.

KEYWORDS
Trustworthy Autonomous Systems; Human-Agent Collectives; Multiagent Responsibility Reasoning; Human-Centred AI

1 INTRODUCTION
To develop and effectively deploy Trustworthy Autonomous Systems (TAS) [33, 65], it is crucial to coordinate their behaviour [69], ensure their compatibility with our human-centred social values [72], and design verifiably safe and reliable human-agent collectives [49]. To that end, we face various social, technological, legal, and ethical challenges for which socio-technically expressive notions of responsibility, blameworthiness, and accountability need to be developed. This requires an interdisciplinary effort as it relates to:

- Philosophy of AI, Applied Ethics, and Ethics by Design: Studying the conceptual links between the notion of autonomy and responsibility in human-agent collectives;
- Sociological Aspects of Agency and Autonomy: Capturing the social implications of the introduction of autonomous systems into society and conceptualising how different levels of autonomy relate to different notions of responsibility;
- Legal Studies and Automated Judicial Reasoning Tools: Formalising legal principles, based on the jurisprudential perspective on responsibility, to govern autonomous systems towards preserving social values and contextual norms;
- Design Methodologies: Integrating value-based and co-active design methods to ensure responsibility in and by design; and
- Multiagent Technologies and Formal Methods: Developing automated responsibility reasoning tools and decision support services for human-centred autonomous systems.

The need for ensuring trustworthiness of autonomous systems is known and well-argued in the literature [30, 59]. However, as long as we remain at an abstract level and merely discuss how TASs ought to behave (i.e., without clear instructions on potential ways to ensure trustworthiness), the gap will not be bridged. We argue that to ensure TAS, the community requires intermediary notions, common languages, and operational tools to represent and reason about different facets of trustworthiness in the context of TAS. We require a notion that is, on one hand, rich-enough to capture the aforementioned (philosophical, social, legal, technological, and design) aspects of TAS and, on the other, computationally implementable (i.e., for which there exist formal models and expressive reasoning tools). To address this gap, we deem that the multidimensional notion of responsibility in its various forms (e.g., blameworthiness, accountability, sanctionability, and liability) can be used, tailored, and extended for this purpose.

With more autonomy comes more and different forms of responsibility.

In principle, responsibility necessitates autonomy as this is defined only for an agent with a level of autonomy [15, 52]. From the other side, autonomy is about the capacity of an entity to manifest its agency via performing actions [74, 75], and causing change in the environment to reach its desires [16, 17, 39, 70]. Then agent A causing change and reaching outcome O in the environment indicates “A’s responsibility for O” [23, 44–46]. We see that an agent’s responsibility can be formulated in terms of the post-conditions of their actions as an ex post notion (i.e., whether the execution of affordable strategies already resulted in an outcome for which agents are responsible). As a complementary approach, the line of research on action-state semantics [18, 71, 90, 92] focuses on the strategic capacities of agents with respect to potential situations...
in prospect. In this view, agents' responsibility is formulated in terms of pre-conditions as an ex ante notion. These two forms of retrospective and prospective responsibility are key for conceptualising what van de Poel [82] calls backward- and forward-looking notions of responsibility. On the other hand, Santoni de Sio and van den Hoven [73] argue that ultimately it should be humans not computers and their algorithms that are to remain in control of, and thus morally responsible for, relevant decisions. This is captured under the notion of "meaningful human control". However, one must realise that humans must be in a position to reason about, and capable of understanding, what part in a system they are expected to 'take over control of' and at which appropriate moment. As designers and engineers of these algorithms, it is in turn our responsibility to ensure that we design our systems in a way that these criteria can be met.

In this proposal, we show how different dimensions of responsibility relate to challenges in development and deployment of TAS. This is the first attempt to articulate TAS challenges to which responsibility reasoning can contribute and is a starting point for establishing a research agenda on "Responsibility Research for Trustworthy Autonomous Systems". This work is structured based on the two categories of prospective and retrospective responsibilities. For both, we elaborate on TAS challenges and open research problems, and present a way forward by sketching methods that we see well-suited to investigating these problems.

2 PROSPECTIVE RESPONSIBILITY IN TAS

Prospective responsibility reasoning is focused on eventualities as situations that may materialise in future and analyses how agents can or ought to affect such state of affairs. In autonomous systems, prospective responsibility reasoning is crucial, e.g., to ascribe the responsibility for ensuring the safety of an autonomous vehicle system to a capable agent or agent group. This calls for considering the strategic abilities of humans as well as artificial entities in responsibility reasoning and, in turn, in assigning tasks to human-agent collectives. Moreover, responsibility reasoning can be of use to design verifiably reliable autonomous human-agent organisations. Below, we present TAS challenges that call for novel responsibility reasoning research and discuss desirable requirements to be met.

**Challenge 1.** The need for practical and provably sound degrees of responsibility to ensure system reliability and fault tolerance in the technical software development context.

In real-life autonomous systems, reliability of the system and its ability to handle potential failures are key for social acceptance. The society will not accept the integration of autonomous vehicles unless they show the capacity to perform reliably and in a fault-tolerant manner. One should never expect that all the components in an autonomous system behave as expected, and so one has to put in place overarching methods to ensure reliability. For this, we can rely on formally verifiable responsibility reasoning methods [61, 92]. Following Chockler and Halpern [23], we deem that the notion of responsibility can be a base for conceptualising resilience. (See Moshe Vardi’s call on the need for methods capable of analysing the trade-off between efficiency and redundancy in socio-technical systems and for developing comprehensive models of resilience [86].) We suggest modelling resilience in TAS in terms of responsibility degrees. Imagine a 3-member multiagent software system in which only agent A has the full responsibility with respect to updating a block/value (task responsibility). It means that if A fails, no-one is able to correct the problem. If the system was designed such that at least two (coordinated) agents had a non-zero degree of responsibility for updating the block/value, we have redundancy but control is distributed. Such a system is more resilient against potential failures. We propose further investigation on how different formalisations of the notion of responsibility (e.g., the causal notion of [23] or the strategic notion of [92]) can be of use in different domains to ensure the resiliency of autonomous systems.

**Challenge 2.** The need for operational accountability ascription and task coordination methods in TAS’s organisational context.

In human-agent collectives, where human and artificial agents collaborate, it is crucial to put in place mechanisms for balancing the two decision-making types in what Jennings et al. call flexible autonomy [49]. In essence, flexible autonomous systems allow "agents to sometimes take actions in a completely autonomous way without reference to humans [type 1], while at other times being guided by much closer human involvement [type 2]". Then the main problem is to understand who is, and to what extent they are, accountable for the outcome of such decisions. A way forward is to employ Multiagent Organisation (MAO) models [34, 47, 73] and develop accountability ascription methods for human-agent autonomous systems. Such methods are expected to be expressive to reason about task coordination, delegation, and shared control in TAS [35, 62, 91] and be dynamic for moving between the two types of decision making. Moreover, to ensure reliability in human-agent organisations, accountability reasoning can be used as a mechanism to provide explanation for outcomes [7, 9].

3 RETROSPECTIVE RESPONSIBILITY IN TAS

Imagine a multiagent system that includes autonomous vehicles, pedestrians, and human-driven vehicle. After the occurrence of a crash, retrospective responsibility is to reason about individuals or groups of agents capable of avoiding the crash (retrospective responsibility in terms of avoidance power [18]) or those who caused it (retrospective responsibility in terms of causal affirmative power [23]). Computational retrospective responsibility tools can be of use for automated liability determination in TAS, for addressing the so-called responsibility gaps/voids (where a group is determined to be responsible collectively but individuals’ share is not clear), and for building sanctioning tools and value-aligned coordination mechanisms to ensure the functionality of TAS.

**Challenge 3.** The need for tools to address responsibility voids in human-agent collectives and measures to fairly distribute collective-level responsibilities into individual-level degrees of responsibility.

Imagine a scenario (adapted from [54]) where a traveller’s water canteen is poisoned by one and then emptied by another fellow traveller. The traveller dies of thirst in the middle of the desert. It is clear that the two fellow travellers are responsible as a collective but the extent of responsibility of each is not clear. This is a stranded case of the so called "responsibility void" [14] where linking collective to individual responsibility is a challenge. In the
responsibility literature, there exist suggestions to adopt cost allocation techniques to ascribe responsibility among agents with respect to their contribution to the collective [37, 92]. While such approaches lead to desirable fairness properties they are not scalable due to their expensive computational complexity. This is more challenging in mixed (human-artificial) teams [49] with flexible autonomy in place. These are collectives in which artificial agents sometimes make decisions with complete autonomy and sometimes operate under more control from humans. For instance, imagine a healthcare scenario where human surgeons are performing an operation in collaboration with semi-autonomous robots. Who is, and to what extent are they responsible for a potential failure? As we are faced with dynamic degrees of autonomy, we require methods that are able to ascribe responsibility dynamically. A way forward is to capture resource and cost dynamics [3, 5] (i.e., who had control over what resource in which time period) for responsibility reasoning; and to integrate methods that consider real-life limitations of goals/tasks to allow tractable ability verification [10, 41].

**Challenge 4.** *The need for context-aware blameworthiness and accountability reasoning tools as a basis for effective liability measures and to ensure the legality of TAS.*

By giving more autonomy to artificial systems, one cannot still see them as object-like tools that merely follow instructions. For instance, a driver-less vehicle is not receiving direct instructions thus, when collisions occur, a judge cannot simply apply "Qui facit per alium, facit per se" (who acts through another does the act himself) [24, 51, 64] to see the owner as the only responsible agent. It is reasonable that any involved agent with a degree of autonomy takes a degree of blameworthiness. However, on a basic level, most of our enforcement methods are founded on physical regimentation techniques, e.g., to imprison or impose some form of physical restriction, that are neither effective on, nor meaningful for non-human agents. We deem that, for effective deployment of autonomous systems, it is neither effective nor efficient to rely on non-automated resource-consuming judiciary processes. By doing so, we are automating transportation and manufacturing but need to add much more capacities (human labour, time, and judiciary expertise) to judge every incident of failure. This is not an attempt for automating the judiciary system but, in contrast, a proposal to capture the capacities of non-human agents, consider social values, and develop human-centred legal decision support tools for TAS. To merge human-dependent enforcement methods (e.g., imposing limitations on resources) with coordination mechanisms that are applicable to artificial agents, the literature on normative multiagent systems [12] offers methods for incentive engineering and norm-aware mechanism design [19, 20], techniques for sanction-based enforcement [27, 87], and models for integrating social norms and ethical values into governance of socio-technical systems [78, 83]. Such methods provide a base for effective liability measures. (As discussed, the retributive perspective on punishment [42] is meaningless for an artificial agent.) This normative approach corresponds with the utilitarian punishment view [6, 11] and the application of criminal deterrence theory [60, 67]. This is to impose sanctions with the goal to nudge the behaviour of autonomous agents, and in turn the behaviour of the collective, towards human-centred values. And in addition, it corresponds with computationally implementable approaches in safe multiagent reinforcement learning [38, 77] where agents’ degree of blameworthiness can be used as a measure of regret or to inform reward-shaping mechanisms.

## 4 Towards a Research Agenda

To investigate how different forms of responsibility reasoning support TAS, we envisage the following research themes (Figure 1 depicts various sub-domains and related research).

**Theme 1.** *Responsibility-aware agents and multiagent systems.*

In general, meta-reasoning refers to the capacity of agents to reflect on their own reasoning [25]. While being able to analyse inputs and flexibly choose an optimal action with respect to the agent’s goals defines it to be intelligent [88], we see responsibility reasoning as a meta-level capacity (on top of self and situation awareness [28, 80]) that enables an agent to be aware of and reason about its own responsibilities and the responsibilities of other human/artificial agents. In this way, a responsibility-aware agent would be able to reason about the consequences of its available actions not only in view of its own goals but also with respect to its degree of responsibility for potential consequences. Following Dignum’s suggested architecture for social agents [31], we envisage responsibility-aware autonomous systems to weigh their options based on operational optimality (e.g., cost efficiency regarding the consumption of energy and time) and in addition have a meta-level responsibility-oriented unit to represent and reason about their degree of responsibility under different eventualities.

**Theme 2.** *Tools for responsibility reasoning under norm conflict.*

Norm conflicts are situation where an agent’s compliance with one norm results in the violation of another. For instance, imagine an autonomous vehicle with a passenger on board who urgently requires medical attention. Through the journey to the hospital, the vehicle is forced to choose between keeping its speed below the safe limit (which increases the chance of arriving late and causing harm to its passengers) or going above the speed limit (which violates safety norms). Both options are normatively undesirable as they violate established norms. As discussed in [13], resolving such situations and understanding how to ascribe responsibilities to the agents involved are crucial for ensuring the reliability of autonomous systems. To address this, we aim to develop norm ranking tools, rooted in argumentation theory [57, 66] and value-aware norm selection methods [76], as a base for formulating novel responsibility degrees that capture a ranked set of norms. (In a future in which the AI technology permeates our society, one can imagine that the knowledge of the predicament and norms is distributed and any agent (partially) aware of the situation can help solve the problem. This calls for investigations on how distributed situation awareness [79] relates to responsibility reasoning.)

**Theme 3.** *Integrated data-driven and model-based responsibility reasoning, and tools for ascribing responsibilities under uncertainty.*

In dynamic multiagent settings, the knowledge agents have about their environment, their abilities, and abilities of others may be imperfect. This also includes their (imperfect) understanding of established norms.¹ In such settings, agents may learn about norms,

¹An agent’s knowledge affects different forms of responsibility differently; e.g., knowledge is crucial for distinguishing blameworthiness from responsibility [23].
and norm changes [21], as the system evolves. To capture such dynamics and model dynamic notions of responsibility, we ideate the integration of methods capable of learning norms and preferences [2, 27, 58, 93] into logic-based frameworks that allow the combination of symbolic and sub-symbolic features of the environment [26, 40, 48]. Such an integration allows reasoning in a probabilistic or possibilistic fashion and formulating hybrid learned-reasoned notions of responsibility.

Acknowledgements. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) platform grant EP/R029563/1 entitled “AutoTrust: Designing a Human-Centred Trusted, Secure, Intelligent and Usable Internet of Vehicles”. The authors also thank the anonymous referees for their valuable comments and helpful suggestions.

Figure 1: Responsibility Reasoning for TAS (Research Avenues and Related Work).