Responsibility Research for Trustworthy Autonomous Systems

Blue Sky Ideas Track

Vahid Yazdanpanah
University of Southampton
Southampton, United Kingdom
v.yazdanpanah@soton.ac.uk

Enrico H. Gerding
University of Southampton
Southampton, United Kingdom
eg@ecs.soton.ac.uk

Sebastian Stein
University of Southampton
Southampton, United Kingdom
ss2@ecs.soton.ac.uk

Mehdi Dastani
Utrecht University
Utrecht, The Netherlands
m.m.dastani@uu.nl

Catholijn M. Jonker
Delft University of Technology
Delft, The Netherlands
c.m.jonker@tudelft.nl

Timothy J. Norman
University of Southampton
Southampton, United Kingdom
t.j.norman@soton.ac.uk

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

ACM Reference Format:

Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, and Timothy J. Norman. 2021. Responsibility Research for Trustworthy Autonomous Systems: Blue Sky Ideas Track. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 6 pages.

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

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

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 Dignums' 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 solving 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,

[1] An agent's knowledge affects different forms of responsibility differently; e.g., knowledge is crucial for distinguishing *blameworthiness* from *responsibility* [23].

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

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.

REFERENCES

[1] Dhaminda B. Abeywickrama, Corina Cîrstea, and Sarvapali D. Ramchurn. 2019. Model Checking Human-Agent Collectives for Responsible AI. In *28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019, New Delhi, India, October 14-18, 2019*. IEEE, New York, NY, 1–8. https://doi.org/10.1109/RO-MAN46459.2019.8956429

[2] Abhijin Adiga, Sarit Kraus, and S. S. Ravi. 2020. Boolean Games: Inferring Agents' Goals Using Taxation Queries. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*. IFAAMAS, Richland, SC, 1735–1737. https://dl.acm.org/doi/abs/10.5555/3398761.3398965

[3] Natasha Alechina, Stéphane Demri, and Brian Logan. 2020. Parameterised Resource-Bounded ATL. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, Palo Alto, CA, 7040–7046. https://aaai.org/ojs/index.php/AAAI/article/view/6189

[4] Natasha Alechina, Joseph Y. Halpern, and Brian Logan. 2017. Causality, Responsibility and Blame in Team Plans. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*. IFAAMAS, Richland, SC, 1091–1099. https://dl.acm.org/doi/10.5555/3091125.3091279

[5] Natasha Alechina and Brian Logan. 2020. State of the Art in Logics for Verification of Resource-Bounded Multi-Agent Systems. In *Fields of Logic and Computation III - Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*. Springer, Cham, 9–29. https://doi.org/10.1007/978-3-030-48006-6_2

[6] Mirko Bagaric. 1999. In defence of a utilitarian theory of punishment: Punishing the innocent and the compatibility of utilitarianism and rights. *Australian Journal of Legal Philosophy* 24 (1999), 95.

[7] Matteo Baldoni, Cristina Baroglio, Olivier Boissier, Roberto Micalizio, and Stefano Tedeschi. 2019. Accountability and Responsibility in Multiagent Organizations for Engineering Business Processes. In *Engineering Multi-Agent Systems - 7th International Workshop, EMAS 2019, Montreal, QC, Canada, May 13-14, 2019, Revised Selected Papers*. Springer, Cham, 3–24. https://doi.org/10.1007/978-3-030-51417-4_1

[8] Matteo Baldoni, Cristina Baroglio, Olivier Boissier, Roberto Micalizio, and Stefano Tedeschi. 2019. Engineering Business Processes through Accountability and Agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*. IFAAMAS, Richland, SC, 1796–1798. https://dl.acm.org/doi/10.5555/3306127.3331922

[9] Matteo Baldoni, Cristina Baroglio, and Roberto Micalizio. 2019. Accountability, responsibility and robustness in agent organizations. In *The 1st International Workshop on Responsible Artificial Intelligence Agents, RAIA 2019*. CEUR-WS.org, Aachen, 1–8.

[10] Francesco Belardinelli, Wojciech Jamroga, Damian Kurpiewski, Vadim Malvone, and Aniello Murano. 2019. Strategy Logic with Simple Goals: Tractable Reasoning about Strategies. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*. ijcai.org, Macao, China, August 10-16, 2019, 88–94. https://doi.org/10.24963/ijcai.2019/13

[11] Jeremy Bentham. 1830. *The rationale of punishment*. R. Heward, London.

[12] Guido Boella, Leendert W. N. van der Torre, and Harko Verhagen. 2006. Introduction to normative multiagent systems. *Computational and Mathematical Organization Theory* 12, 2-3 (2006), 71–79. https://doi.org/10.1007/s10588-006-9537-7

[13] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.

[14] Matthew Braham and Martin van Hees. 2011. Responsibility voids. *The Philosophical Quarterly* 61, 242 (2011), 6–15.

[15] Matthew Braham and Martin van Hees. 2012. An anatomy of moral responsibility. *Mind* 121, 483 (2012), 601–634.

[16] Michael E Bratman. 2007. *Structures of agency: Essays*. Oxford University Press, Oxford.

[17] Michael E. Bratman, David J. Israel, and Martha E. Pollack. 1988. Plans and resource-bounded practical reasoning. *Computational Intelligence* 4 (1988), 349–355. https://doi.org/10.1111/j.1467-8640.1988.tb00284.x

[18] Nils Bulling and Mehdi Dastani. 2013. Coalitional Responsibility in Strategic Settings. In *Proceedings of the 14th International Workshop on Computational Logic in Multi-Agent Systems, CLIMA XIV, Corunna, Spain, September 16-18*. Springer, Berlin, Heidelberg, 172–189. https://doi.org/10.1007/978-3-642-40624-9_11

[19] Nils Bulling and Mehdi Dastani. 2016. Norm-based mechanism design. *Artificial Intelligence* 239 (2016), 97–142. https://doi.org/10.1016/j.artint.2016.07.001

[20] Cristiano Castelfranchi. 1998. Modelling Social Action for AI Agents. *Artificial Intelligence* 103, 1-2 (1998), 157–182. https://doi.org/10.1016/S0004-3702(98)00056-3

[21] Cristiano Castelfranchi. 2015. A Cognitive Framing for Norm Change. In *Proceedings of the 11th International Workshop on Coordination, Organizations, Institutions, and Normes in Agent Systems - COIN 2015, Istanbul, Turkey, May 4, 2015*. Springer, Cham, 22–41. https://doi.org/10.1007/978-3-319-42691-4_2

[22] Cristiano Castelfranchi, Frank Dignum, Catholijn M. Jonker, and Jan Treur. 1999. Deliberative Normative Agents: Principles and Architecture. In *Proceedings of the 6th International Workshop on Intelligent Agents, Agent Theories, Architectures, and Languages, ATAL '99, Orlando, Florida, USA, July 15-17*. Springer, Berlin, Heidelberg, 364–378. https://doi.org/10.1007/10719619_27

[23] Hana Chockler and Joseph Y. Halpern. 2004. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research* 22 (2004), 93–115. https://doi.org/10.1613/jair.1391

[24] Alfred Conard. 1948. What's Wrong With Agency. *Journal of Legal Education* 1 (1948), 540.

[25] Michael T. Cox and Anita Raja. 2011. *Metareasoning: Thinking about thinking*. MIT Press, Cambridge, MA.

[26] Davide Dell'Anna, Fabiano Dalpiaz, and Mehdi Dastani. 2019. Requirements-driven evolution of sociotechnical systems via probabilistic reasoning and hill climbing. *Automated Software Engineering* 26, 3 (2019), 513–557. https://doi.org/10.1007/s10515-019-00255-5

[27] Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. 2020. Runtime revision of sanctions in normative multi-agent systems. *Autonomous Agents and Multi Agent Systems* 34, 2 (2020), 43. https://doi.org/10.1007/s10458-020-09465-8

[28] Louise A. Dennis and Michael Fisher. 2020. Verifiable Self-Aware Agent-Based Autonomous Systems. *Proc. IEEE* 108, 7 (2020), 1011–1026. https://doi.org/10.1109/JPROC.2020.2991262

[29] Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1–14. https://doi.org/10.1016/j.robot.2015.11.012

[30] Virginia Dignum. 2019. *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way*. Springer, Cham. https://doi.org/10.1007/978-3-030-30371-6

[31] Virginia Dignum and Frank Dignum. 2020. Agents are Dead. Long live Agents!. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*. IFAAMAS, Richland, SC, 1701–1705. https://dl.acm.org/doi/abs/10.5555/3398761.3398957

[32] Hein Duijf. 2018. Responsibility voids and cooperation. *Philosophy of the social sciences* 48, 4 (2018), 434–460. https://doi.org/10.1177/0048393118767084

[33] European Commission: The High-Level Expert Group on AI. 2019. Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed: 2021-02-15.

[34] Jacques Ferber, Olivier Gutknecht, and Fabien Michel. 2003. From Agents to Organizations: An Organizational View of Multi-agent Systems. In *Proceedings of the 4th International Workshop on Agent-Oriented Software Engineering, AOSE 2003, Melbourne, Australia, July 15, 2003*. Springer, Berlin, Heidelberg, 214–230. https://doi.org/10.1007/978-3-540-24620-6_15

[35] Frank Flemisch, David A. Abbink, Makoto Itoh, Marie-Pierre Pacaux-Lemoine, and Gina Weßel. 2016. Shared control is the sharp end of cooperation: Towards a common framework of joint action, shared control and human machine cooperation. *IFAC-PapersOnLine* 49, 19 (2016), 72–77. https://doi.org/10.1016/j.ifacol.2016.10.464

[36] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (2019), 261–262. https://doi.org/10.1038/s42256-019-0055-y

[37] Meir Friedenberg and Joseph Y. Halpern. 2019. Blameworthiness in Multi-Agent Settings. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, Palo Alto, CA, 525–532. https://doi.org/10.1609/aaai.v33i01.3301525

[38] Luca Gasparini, Timothy J. Norman, and Martin J. Kollingbaum. 2018. Severity-sensitive norm-governed multi-agent planning. *Autonomous Agents and Multi Agent Systems* 32, 1 (2018), 26–58. https://doi.org/10.1007/s10458-017-9372-x

[39] Michael P. Georgeff, Barney Pell, Martha E. Pollack, Milind Tambe, and Michael J. Wooldridge. 1998. The Belief-Desire-Intention Model of Agency. In *Intelligent Agents V, Agent Theories, Architectures, and Languages, 5th International Workshop, ATAL '98, Paris, France, July 4-7, 1998, Proceedings*. Springer, Berlin, Heidelberg, 1–10. https://doi.org/10.1007/3-540-49057-4_1

[40] Giuseppe De Giacomo, Luca Iocchi, Marco Favorito, and Fabio Patrizi. 2020. Restraining Bolts for Reinforcement Learning Agents. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, Palo Alto, CA, 13659–13662. https://aaai.org/ojs/index.php/AAAI/article/view/7114

[41] Giuseppe De Giacomo and Moshe Y. Vardi. 2015. Synthesis for LTL and LDL on Finite Traces. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*. ijcai.org, Buenos Aires, Argentina, July 25-31, 2015, 1558–1564. http://ijcai.org/Abstract/15/223

[42] Jack P. Gibbs. 1978. Death Penalty, Retribution and Penal Policy, The. *Criminal Law & Criminology* 69 (1978), 291.

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

[44] Joseph Y. Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science* 56, 4 (2005), 843–887.

[45] Joseph Y. Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of*

science 56, 4 (2005), 889–911.
[46] Charles L. Hamblin. 1987. Imperatives.
[47] Bryan Horling and Victor R. Lesser. 2004. A survey of multi-agent organizational paradigms. *Knowledge Engineering Review* 19, 4 (2004), 281–316. https://doi.org/10.1017/S0269888905000317
[48] Hykel Hosni and Enrico Marchioni. 2019. Possibilistic randomisation in strategic-form games. *International Journal of Approximate Reasoning* 114 (2019), 204–225. https://doi.org/10.1016/j.ijar.2019.08.008
[49] Nicholas R. Jennings, Luc Moreau, David Nicholson, Sarvapali D. Ramchurn, Stephen J. Roberts, Tom Rodden, and Alex Rogers. 2014. Human-agent collectives. *Commun. ACM* 57, 12 (2014), 80–88. https://doi.org/10.1145/2629559
[50] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna van Riemsdijk, and Maarten Sierhuis. 2014. Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction* 3, 1 (2014), 43–69. https://doi.org/10.5898/JHRI.3.1.Johnson
[51] Patrick Lin, Keith Abney, and George A. Bekey. 2012. *Robot ethics: the ethical and social implications of robotics.* MIT Press, Cambridge, MA.
[52] John R. Lucas. 1995. *Responsibility.* Clarendon Press, Oxford.
[53] Michael Luck, Samhar Mahmoud, Felipe Meneguzzi, Martin Kollingbaum, Timothy J. Norman, Natalia Criado, and Moser Silva Fagundes. 2013. *Normative Agents.* Springer, Dordrecht, 209–220. https://doi.org/10.1007/978-94-007-5583-3_14
[54] James A. McLaughlin. 1925. Proximate cause. *Harvard Law Review* 39, 2 (1925), 149–199.
[55] Rijk Mercuur, Virginia Dignum, and Catholijn M. Jonker. 2019. The Value of Values and Norms in Social Simulation. *Journal of Artificial Societies and Social Simulation* 22, 1 (2019), 9. https://doi.org/10.18564/jasss.3929
[56] Simon Miles, Steve Munroe, Michael Luck, and Luc Moreau. 2007. Modelling the provenance of data in autonomous systems. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007), Honolulu, Hawaii, USA, May 14-18, 2007.* IFAAMAS, Richland, SC, 50. https://doi.org/10.1145/1329125.1329185
[57] Sanjay Modgil and Michael Luck. 2008. Argumentation Based Resolution of Conflicts between Desires and Normative Goals. In *Argumentation in Multi-Agent Systems, Fifth International Workshop, ArgMAS 2008, Estoril, Portugal, May 12, 2008. Revised Selected and Invited Papers*, Vol. 5384. Springer, Berlin, Heidelberg, 19–36. https://doi.org/10.1007/978-3-642-00207-6_2
[58] Martin Mozina, Jure Zabkar, and Ivan Bratko. 2007. Argument based machine learning. *Artificial Intelligence* 171, 10-15 (2007), 922–937. https://doi.org/10.1016/j.artint.2007.04.007
[59] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020.* IFAAMAS, Richland, SC, 1706–1710. https://dl.acm.org/doi/abs/10.5555/3398761.3398958
[60] Ilene H. Nagel. 1989. Structuring sentencing discretion: The new federal sentencing guidelines. *Journal of Crimiminal Law & Criminology* 80 (1989), 883.
[61] Pavel Naumov and Jia Tao. 2020. An epistemic logic of blameworthiness. *Artificial Intelligence* 283 (2020), 103269. https://doi.org/10.1016/j.artint.2020.103269
[62] Timothy J. Norman and Chris Reed. 2000. Delegation and Responsibility. In *Intelligent Agents VII. Agent Theories Architectures and Languages, 7th International Workshop, ATAL 2000, Boston, MA, USA, July 7-9, 2000, Proceedings*, Vol. 1986. Springer, Berlin, Heidelberg, 136–149. https://doi.org/10.1007/3-540-44631-1_10
[63] Timothy J. Norman and Chris Reed. 2002. Group delegation and responsibility. In *The First International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002, July 15-19, 2002, Bologna, Italy, Proceedings.* ACM, New York, NY, USA, 491–498. https://doi.org/10.1145/544741.544856
[64] Timothy J. Norman and Chris Reed. 2010. A logic of delegation. *Artificial Intelligence* 174, 1 (2010), 51–71. https://doi.org/10.1016/j.artint.2009.10.001
[65] Office for Artificial Intelligence. 2020. A guide to using artificial intelligence in the public sector. https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector. Accessed: 2021-02-15.
[66] Nir Oren, Michael Luck, Simon Miles, and Timothy J. Norman. 2008. An argumentation inspired heuristic for resolving normative conflict. In Proceedings of the Fifth Workshop on Coordination, Organizations, Institutions and Norms in Agent Systems, COIN@AAMAS 2008.
[67] Samuel H Pillsbury. 1989. Understanding penal reform: The dynamic of change. *Journal of Criminal Law & Criminology* 80 (1989), 726.
[68] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
[69] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486. https://doi.org/10.1038/s41586-019-1138-y
[70] Anand S. Rao and Michael Wooldridge. 1999. *Foundations of Rational Agency.* Springer, Dordrecht, 1–10. https://doi.org/10.1007/978-94-015-9204-8_1
[71] Chris Reed and Timothy J. Norman. 2007. A Formal Characterisation of Hamblin's Action-State Semantics. *Journal of Philosophical Logic* 36, 4 (2007), 415–448. https://doi.org/10.1007/s10992-006-9041-z
[72] Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control.* Viking, New York, NY.
[73] Filippo Santoni de Sio and Jeroen van den Hoven. 2018. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers Robotics AI* 5 (2018), 15. https://doi.org/10.3389/frobt.2018.00015
[74] John R. Searle. 1989. How performatives work. *Linguistics and philosophy* 12, 5 (1989), 535–558.
[75] John R Searle. 1995. *The construction of social reality.* Free Press, New York, NY.
[76] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Manel Rodríguez, Michael J. Wooldridge, Javier Morales, and Carlos Ansótegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018.* IFAAMAS, Richland, SC, 1294–1302. http://dl.acm.org/citation.cfm?id=3237891
[77] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. *CoRR* abs/1610.03295 (2016), 13. http://arxiv.org/abs/1610.03295
[78] Munindar P. Singh. 2013. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology* 5, 1 (2013), 21:1–21:23. https://doi.org/10.1145/2542182.2542203
[79] Neville A. Stanton. 2016. Distributed situation awareness. *Theoretical Issues in Ergonomics Science* 17, 1 (2016), 1–7. https://doi.org/10.1080/1463922X.2015.1106615
[80] Neville A. Stanton, Paul M. Salmon, Guy H. Walker, Eduardo Salas, and Peter A. Hancock. 2017. State-of-science: situation awareness in individuals, teams and systems. *Ergonomics* 60, 4 (2017), 449–466. https://doi.org/10.1080/00140139.2017.1278796
[81] Nicolas Troquard. 2014. Reasoning about coalitional agency and ability in the logics of "bringing-it-about". *Autonomous Agents and Multi Agent Systems* 28, 3 (2014), 381–407. https://doi.org/10.1007/s10458-013-9229-x
[82] Ibo van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral responsibility.* Springer, Dordrecht, 37–52. https://doi.org/10.1007/978-94-007-1878-4_3
[83] Ibo van de Poel. 2020. Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines* 30 (2020), 1–25. https://doi.org/10.1007/s11023-020-09537-4
[84] Ibo van de Poel and Lambèr Royakkers. 2011. *Ethics, technology, and engineering: An introduction.* John Wiley & Sons, Hoboken, NJ.
[85] Jeroen van den Hoven. 2013. *Value Sensitive Design and Responsible Innovation.* John Wiley & Sons, Hoboken, NJ, 75–83. https://doi.org/10.1002/9781118551424.ch4
[86] Moshe Y. Vardi. 2020. Efficiency vs. resilience: what COVID-19 teaches computing. *Commun. ACM* 63, 5 (2020), 9. https://doi.org/10.1145/3388890
[87] Daniel Villatoro, Giulia Andrighetto, Jordi Sabater-Mir, and Rosaria Conte. 2011. Dynamic Sanctioning for Robust and Cost-Efficient Norm Compliance. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence.* ijcai.org, Barcelona, Catalonia, Spain, July 16-22, 2011, 414–419. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-077
[88] Michael J. Wooldridge and Nicholas R. Jennings. 1995. Intelligent agents: theory and practice. *The knowledge engineering review* 10, 2 (1995), 115–152. https://doi.org/10.1017/S0269888900008122
[89] Vahid Yazdanpanah and Mehdi Dastani. 2015. Quantified Degrees of Group Responsibility. In *Coordination, Organizations, Institutions, and Normes in Agent Systems XI - COIN 2015 International Workshops, COIN@AAMAS, Istanbul, Turkey, May 4, 2015*, Vol. 9628. Springer, Cham, 418–436. https://doi.org/10.1007/978-3-319-42691-4_23
[90] Vahid Yazdanpanah and Mehdi Dastani. 2016. Distant Group Responsibility in Multi-agent Systems. In *PRIMA 2016: Princiles and Practice of Multi-Agent Systems - 19th International Conference, Phuket, Thailand, August 22-26, 2016, Proceedings.* Springer, Cham, 261–278. https://doi.org/10.1007/978-3-319-44832-9_16
[91] Vahid Yazdanpanah, Mehdi Dastani, Shaheen Fatima, Nicholas R. Jennings, Devrim Murat Yazan, and W. Henk M. Zijm. 2020. Multiagent Task Coordination as Task Allocation Plus Task Responsibility. In *Multi-Agent Systems and Agreement Technologies - 17th European Conference, EUMAS 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers*, Vol. 12520. Springer, Cham, 571–588. https://doi.org/10.1007/978-3-030-66412-1_37
[92] Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina, and Brian Logan. 2019. Strategic Responsibility Under Imperfect Information. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019.* IFAAMAS, Richland, SC, 592–600. http://dl.acm.org/citation.cfm?id=3331745
[93] Yijie Zhang, Roxana Radulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2020. Opponent Modelling for Reinforcement Learning in Multi-Objective Normal Form Games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020.* IFAAMAS, Richland, SC, 2080–2082. https://dl.acm.org/doi/abs/10.5555/3398761.3399081