Citizen-Centric Multiagent Systems

Blue Sky Ideas Track

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

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

ABSTRACT

Advances in multiagent systems (MAS) have the potential to solve critical societal challenges. For example, MAS techniques for efficient resource allocation can help us implement cleaner and more efficient forms of on-demand mobility; social choice methods can support us in deciding how to trade off energy use and comfort in smart buildings; and task coordination methods can be used to respond to disasters in an effective and resilient manner. However, the benefits of these approaches can only be realised if citizen end users are able to trust these emerging multiagent systems. To achieve this, a *citizen-centric* approach needs to be taken. This places citizens at the heart of the design, development and deployment of trustworthy multiagent systems. We present open research challenges in this area, put forward key application domains for citizen-centric MAS (C-MAS) and discuss collaborative research opportunities.

KEYWORDS

Citizen-Centric Multiagent Systems; AI for Social Good

ACM Reference Format:
Sebastian Stein and Vahid Yazdanpanah. 2023. Citizen-Centric Multiagent Systems: Blue Sky Ideas Track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 6 pages.

1 INTRODUCTION

AI systems are increasingly used to support and often automate decision-making on an unprecedented scale. Such AI systems can draw on a vast range of data sources to make fast, efficient, data-driven decisions to address important societal challenges and potentially benefit millions of citizens [43, 58]. Key application areas include the management of critical infrastructure, including electricity networks [36] and transportation systems [11], or the provision of social services, including policing [26], emergency response [34] and medical support during epidemics like COVID-19 [19].

Most large-scale AI systems, including all the examples given above, are highly distributed multiagent systems. They are characterised by the presence of multiple stakeholders, including autonomous intelligent agents, service providers and citizen end users, that need to make efficient collective decisions despite sometimes conflicting interests. There is a wealth of research in the area of multiagent systems that has considered these types of settings. Specifically, work on game theory, negotiation and mechanism design has looked at how to model self-interested agents, how they can reach agreements or how socially beneficial behaviours can be incentivised [30, 32]. Computational social choice investigates how to derive consensus when making collective decisions [6]. Work on coordination considers how decentralised collectives of agents can collaborate together effectively [56, 61]. However, in this existing work, the citizen end user is usually given a peripheral role, acting as a passive data source or is assumed to be a simple rational decision maker. This ignores the key challenge of ensuring that large-scale AI systems are seen as trustworthy by citizens, an important feature that is needed for the widespread acceptance of safe, reliable and trustworthy AI [17, 35].

More specifically, large-scale AI systems may need access to relevant information from individuals, e.g., their electricity demand, travel preferences or health data, in order to allocate limited resources or services to those that need them most. This raises privacy issues and may also encourage strategic manipulation, where individuals misrepresent their preferences for personal benefit [6, 16]. Furthermore, the systems must be trusted to act in a manner that aligns with society's ethical values [28]. This includes the minimisation of discrimination and the need to govern such systems and ensure equitable decisions [54]. Finally, there is a need to explain decisions and decision-making processes to non-expert end users and other stakeholders.

In order to address these challenges and design trustworthy multiagent systems that can realise their potential of positively affecting people on a large scale, we argue that it is imperative to take a *citizen-centric* approach.[1] In this approach, citizens are viewed as first-class agents at the centre of these multiagent systems. Citizens' preferences are learnt and modelled explicitly while safeguarding their privacy, the system acts to maximise their utility while ensuring equitable and fair outcomes, and automated decisions are explained clearly and can be audited by all stakeholders.

In the following, we first outline our high-level vision of citizen-centric multiagent systems (Section 2), then we highlight open research challenges (Section 3). This is followed by a number of research opportunities that were co-created with a group of industrial, academic and government stakeholders (Section 4). Section 5 concludes this paper.

2 VISION OF CITIZEN-CENTRIC MAS

Figure 1 summarises our high-level vision of a citizen-centric multiagent system (C-MAS). A key aspect of such a system is that information and control are highly distributed, thus preserving privacy and autonomy for all stakeholders. Specifically, we envisage

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

[1] Here, we use *citizen* to denote an end user of an AI system, or someone directly affected by it. We use this term to include a broad spectrum of users, including non-experts, but also to highlight the democratic nature of the AI systems that we envisage.

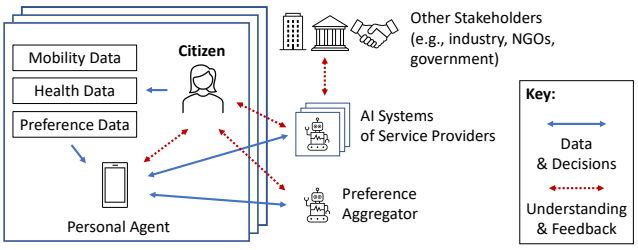

Figure 1: A Citizen-Centric Multiagent System

that citizens will be in control of their private data (e.g., mobility and health data), which will be stored locally on a smart device (e.g., as part of a personal data store [48]) or held securely by a trusted third party. Moreover, each citizen will have a personal intelligent agent that can access this information, learns the citizen's personal preferences and then interacts with external service providers on the citizen's behalf. In smart mobility, this may take the form of arranging an autonomous taxi to take the citizen to work in the morning; in smart energy, the agent might control the citizen's heating and appliances to coincide with the production of cheap renewable energy; and in disaster response, the agent may request tailored help from emergency services.

In these systems, service providers implement their own AI algorithms, e.g., to allocate on-demand mobility services, to offer price incentives to energy consumers or to dispatch emergency responders. Critically, these algorithms will not have full access to each citizen's private information, but will receive at most limited data relevant to a given transaction and in accordance with each citizen's privacy preferences. To help the personal intelligent agents make better decisions, there might be limited exchange of information between different citizens. This would help address the cold-start problem, where new agents with little prior knowledge of their citizens can draw on larger population statistics to initialise their decision-making. Such information exchange would be facilitated by third-party preference aggregators that anonymise and share data according to transparent and customisable rules, similar to data trusts [22, 31]. Finally, a key aspect of C-MAS is that citizens and other stakeholders are able to understand how the AI algorithms work and make decisions.[2]

3 RESEARCH CHALLENGES FOR C-MAS

To ensure that our vision of C-MAS is realised and that citizens are able to trust these systems, they need to be:

- **Citizen-aware:** being aware of the preferences, needs and constraints of individuals, in order to provide personalised and appropriate services, while also respecting privacy constraints.
- **Citizen-beneficial:** acting to maximise the utility of citizens, including the provision of incentives to encourage socially-beneficial behaviour changes.
- **Citizen-sensitive:** making fair, inclusive and equitable decisions.
- **Citizen-auditable:** providing explanations for decisions, thus allowing stakeholders to engage in a continuous feedback loop.

In the following, we discuss these features in more detail, highlighting open research challenges to achieve them.

3.1 Achieving Citizen-Aware C-MAS

Citizen-aware multiagent systems are designed to learn the preferences and requirements of different citizens. This is important because it allows personalised services to be provided to citizens, matching their individual needs and preferences, rather than assuming that all citizens have the same preferences.

In C-MAS, this is primarily achieved through personal intelligent agents, each of which interacts with and represents one citizen. Since this departs from typical centralised AI systems that assume availability of rich data sets, a key challenge here is to quickly learn an accurate model of a citizen's preferences given sparse data and without placing excessive cognitive burden on the user. Achieving this is an open research challenge, but could involve techniques like inverse reinforcement learning to infer utility models from observations [1], the use of domain-specific discrete choice models [47] and targeted preference elicitation techniques [3, 42].

Another open challenge is how to allow the selective exchange of information between different personal agents. This would help to identify common patterns and provide suitable priors for the preference models, which are then refined through subsequent observations or queries. This sharing of data could be done directly between trusted agents, or via third-party preference aggregators that collect anonymised data. Addressing this research challenge could draw on work on recommender systems [41], differential privacy [12] and federated learning [24].

Once the preferences and requirements of a citizen have been learnt or elicited, the personal intelligent agent can now represent the citizen within the wider C-MAS and especially in interactions with service providers. Here, the agent acts primarily on behalf of and to the benefit of its owner. It may selectively reveal information when this is in its owner's interest, but only according to the privacy preferences of its owner and with meaningful consent [13, 18]. More broadly, C-MAS should be designed in such a way as to minimise the risk of privacy breaches [52]. This might include the use of encryption and other security measures to protect the data collected from citizens, and robust procedures for managing this data [60].

3.2 Achieving Citizen-Beneficial C-MAS

Following Russell [39] and the argument that autonomous systems should be inherently human-beneficial, we argue for a transition from the citizen-ignorant notion of artificially intelligent agents and multiagent systems to a citizen-centric approach whose main purpose is to benefit citizens. Here, citizen-beneficial multiagent systems are AI-based systems that are designed to provide benefits to society (including citizens, the economy and the environment). These systems can help to address some of the biggest challenges facing society, such as climate change, pollution and inequality.

[2] C-MAS can be seen as a subclass of human-centred AI systems [44, 55], with a focus on (i) engagement and involvement of non-expert users; (ii) bidirectional relations and feedback loops; and (iii) dynamic involvement of end-users. C-MAS goes beyond the standard human-centred approach by focusing on complex multiagent settings that include non-expert users whose preferences have to be captured and reconciled dynamically, while also providing transparent feedback. This makes C-MAS an innovative and promising approach to AI with the potential to create a more inclusive and accessible future for all users and the prospect of trustworthy human-AI partnerships [35, 45].

Achieving citizen-awareness, as covered in the previous section, helps us to also achieve widespread benefits for citizens in a C-MAS. Specifically, once individual preferences have been learnt, an AI-based resource allocation mechanism (e.g., associated with a service provider) could then aggregate these preferences in a way that preserves diversity and inclusivity. This means that the mechanism will not simply choose the most preferred option, or the option that is preferred by the majority of citizens. Instead, using methods rooted in computational social choice theory (e.g., [46]), it will take into account the preferences of all citizens and seek to find a solution that is fair and equitable for all. However, citizens may not always wish to provide their preference information to enable such aggregation mechanisms. Here, building on work in mechanism design and behavioural economics, it becomes important to consider techniques from incentive engineering [35]. This approach allows for the creation of financial incentives that encourage socially beneficial behaviours, such as greener practices or lower resource consumption (see e.g., [37, 58]). This can be done through the introduction of macro-/micro-level subsidies for environmentally-friendly practices and the taxation of historically environmentally-unfriendly practices. By providing these incentives, it becomes financially viable for individuals (and businesses) to adopt greener practices, which can help to reduce the negative impact of human activities on the environment.

A key aspect of incentive engineering in citizen-beneficial multiagent systems is that it must be diversity-aware as different citizens may require different types and levels of incentives in order to change their behaviour. For example, some may be more receptive to financial incentives, while others may be more motivated by non-financial incentives, such as increased social status or recognition. Finally, the process of incentivising should be sustainability-aware, with incentives being provided in view of the social good. The incentives should not only be effective at encouraging positive behaviour change, but they should also be sustainable over the long term.[3] This could involve providing incentives that encourage the development of new technologies or practices that can help to reduce the negative impact of human activities on the environment, while also providing benefits to society as a whole. By adopting these sustainability-aware incentives (e.g., in providing mobility services [4, 25]), citizen-beneficial multiagent systems can help to create a more sustainable future for all citizens.

3.3 Achieving Citizen-Sensitive C-MAS

Citizen-sensitive C-MAS are distributed AI systems that are designed to make fair and equitable decisions in collaboration with humans, and, with the appropriate permissions and consent, on their behalf (e.g., a smart thermostat or navigation application operating on a network of vehicles). These systems are designed to be aware of the context in which they are operating and to take into account the varying perceptions of fairness and equity that different individuals and groups may have. Similarly, this may depend on the domain in which they are being applied.

To achieve this level of citizen-sensitivity and ensure *responsible autonomy* [9, 57], it is necessary to co-define quantitative metrics for equitability and fairness in different domains. These metrics need to be context-aware, as different citizens may have different perceptions of fairness and equity depending on the specific context in which they are being applied. For example, in the healthcare domain, citizens may perceive fixed prices for services as equitable, while in the transportation domain, they may be willing to accept variations in prices that fluctuate with supply.

To that end, one of the key challenges in the design of C-MAS is how to dynamically price AI-assisted (smart) services, e.g., smart mobility and energy management services, in view of equitability and fairness measures. This is an open challenge that AI-based tools can contribute to, but it will require input from a variety of disciplines, including social and behavioural sciences, to develop context-specific metrics for fairness and equity [10]. Another discipline from which C-MAS can benefit is law and legal reasoning for, and in presence of, AI systems [38]. As AI-assisted systems become more widespread, it will be important to capture existing regulatory measures and to develop new ones that are citizen-aware and protective of citizens' rights. Citizen-sensitive multiagent systems can help to support the development of legal decision-making tools and regulatory measures that protect citizens during the design and operation of AI systems, while also avoiding harmful consequences.

3.4 Achieving Citizen-Auditable C-MAS

The development of AI systems that are capable of providing explanations to non-expert citizens is a crucial step towards achieving citizen-auditable multiagent systems. This is because it enables users to understand the reasoning behind the decisions made by these systems, which is necessary for them to be able to monitor and fine-tune their behaviour.[4] This means that the explanations should not just be technical annotations on the inputs and outputs of a particular component, but should also provide insight into the main purpose and aim of the AI system. In addition to providing understandable explanations, an auditable C-MAS should also enable all stakeholders to engage in a continuous feedback loop to adapt the overall AI system to suit their ethical preferences [29]. This allows users to monitor and maintain the ethical behaviour of AI systems in accordance with their values and ethical principles.

By enabling citizens to monitor and fine-tune the behaviour of AI systems with respect to their own ethical preferences, C-MAS can help to transition from a focus on hard, regulated ethics (as rules set by authorities to ensure safety requirements) to a more flexible approach that incorporates "soft ethics" [14, 15]. This allows citizen-centric *ethical governance* [53] by enabling greater adaptability in the ethical behaviour of AI systems and can help to ensure that these systems align with the personal values of their users [23, 29] and wider stakeholders [7]. This will determine how an AI system is expected to behave (set by users) within the legally allowed spectrum (set by authorities).

The development of auditable C-MAS is an important step towards enabling citizens to monitor and maintain the ethical behaviour of AI systems. By providing explanations and allowing

[3] The term "sustainability" typically refers to the ability of a system or solution to endure over time while minimising negative impacts on society, finances and the environment. This concept is often framed by the three pillars of social, financial and environmental sustainability [33]. We believe that C-MAS and the solutions it provides can only be considered practical if they capture all three pillars of sustainability.

[4] Additionally, such explanations can feed into computational methods for monitoring and ascribing responsibility for any harm or unanticipated failures in C-MAS [8].

users to personalise the AI system to suit their own preferences, C-MAS can help ensure that AI systems are trustworthy and safe.

4 C-MAS FOR SOCIAL GOOD

In September 2022, we organised a workshop to validate the concept of C-MAS for achieving social good and to identify concrete research opportunities. The workshop included a diverse group of stakeholders from 15 different organisations, including key industry representatives, academics and government agencies. In this section, we provide a brief overview of the discussions at this workshop, both to illustrate how the concept of C-MAS can be applied in real-world settings for social good, and to highlight concrete research opportunities for academia, industry and policymakers.[5]

4.1 C-MAS Use Cases

At the workshop, we identified and discussed a number of promising use cases for C-MAS, including:

- **Clean Transportation:** By considering the preferences and constraints of individuals, C-MAS can help people switch to cleaner on-demand and shared mobility [21, 25]. This will involve suggesting appropriate modes of transport and reacting to incentives where appropriate (e.g., to delay a journey or switch modes). C-MAS can also help people transition to electric vehicles and deal with a currently limited rapid charging infrastructure by suggesting personalised charging stops on long routes [59].
- **Smart Energy:** Similarly, C-MAS can help citizens optimise their energy at home, heating the house or running appliances when cheap renewable energy is available [2]. Home energy storage, including electric vehicles, can be used to store or even trade energy with the grid or a neighbourhood.
- **Audio AI Services:** Audio AI offers an opportunity to make C-MAS more accessible and seamless for citizen users [51]. Audio services could allow users to interact with their personal intelligent agents and could even collect contextual information about a user's activities or intents (e.g., for adjusting the heating or booking imminent transport). Clearly, this poses additional privacy, trust and safety challenges that need to be addressed [52].
- **Social Recommender Systems:** Since C-MAS are highly distributed with many service providers and users, existing work on trust and reputation systems can be applied to help citizens use information from trusted social contacts to find services [5, 40].

4.2 Collaborative Research Opportunities

Based on the example use cases above, we identified concrete research opportunities with the workshop participants. These are specific steps that academia, industry, representatives of citizens and policymakers can take together to start addressing the research challenges outlined in Section 3.

From Explainability to Transparency: Software developers are an important part of a C-MAS, and there are opportunities for changing the way that citizen-centric AI software is developed and moving towards explainable approaches [20, 27, 50]. Currently, most documentation is targeted towards other developers, but to achieve explainability for product owners (and ultimately transparency for end users), additional annotations for a wider audience may be needed. Such increased transparency will help establish trust in the users of AI systems. Here, it is also important to consider the context in which AI is used, which may affect trust and required accuracy.

Research Opportunity 1. *Investigating the conceptual differences and relations between function-oriented (encapsulated) explainability and purpose-oriented (contextual) transparency. This will be a first step to design and develop AI systems that are transparent to end users (by providing insights about their performance in a particular context) and also explainable for software developers (by providing explanations on their robustness, accuracy and technical reliability).*

From Robustness to Resilience: C-MAS need to be able to deal with potential failures. Here, users have different levels of tolerance towards failures, not only because of inherent heterogeneity among users, but also with respect to the application domain or context. For instance, one may tolerate a minor failure from a mobile mapping app when looking for a particular restaurant, but not in healthcare services or in AI systems that manage critical infrastructure. Thus, building systems that are not just robust to anticipated failures, but also resilient to unforeseen circumstances [49] is important.

Research Opportunity 2. *Distinguishing essential from desirable features in C-MAS in different application domains and identifying key features for context-dependent resilience. This will be an input for shifting from robustness (as an approach for ensuring fault-tolerance) to establishing context-aware resilience in C-MAS (for ensuring that the system can recover from unforeseen circumstances).*

From Reliability to Trustworthiness: Biases in data and unintended harm from AI are a significant danger. There is a need to train and audit AI developers, to monitor AI providers, and to provide effective tools to control malicious actions by AI systems (e.g., to spread misinformation or promote harmful behaviours on social media). It is important here to be aware of who is collecting data and whether we (as a society, individual citizen or a team of citizen representatives) trust them. This is challenging, as trustworthiness is a dynamic notion that needs to be calibrated.

Research Opportunity 3. *Building interdisciplinary approaches to study the applicability of different AI auditing tools and standardisation schemes. Interacting with experts from social science and legal experts to support the development of effective methods for establishment and dynamic calibration of trustworthy C-MAS.*

5 CONCLUSIONS

In this paper, we outlined a vision of citizen-centric multiagent systems, a paradigm for building complex AI systems that can address important societal challenges, but that treat the citizen end users as first-class agents. This involves being aware of the citizens' preferences, taking actions and offering incentives that benefit citizens (and the wider good of society), being sensitive to all members of society and engaging all stakeholders in a continuous feedback loop. Achieving such citizen-centric multiagent systems will help us build trustworthy and widely accepted AI systems that can improve our quality of life, support us in addressing the climate emergency and make our society more resilient.

[5] In addition to this workshop, and through various engagements and outreach-oriented activities, we presented the C-MAS perspective to non-expert end users and received feedback on the expectations of a diverse and inclusive range of citizens.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems (https://ccais.ac.uk/) and through the AutoTrust Platform Grant (EP/R029563/1).

REFERENCES

[1] Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297 (2021), 103500. https://doi.org/10.1016/j.artint.2021.103500

[2] Frederik Auffenberg, Stephen Snow, Sebastian Stein, and Alex Rogers. 2017. A comfort-based approach to smart heating and air conditioning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 3 (2017), 1–20.

[3] Tim Baarslag and Enrico H. Gerding. 2015. Optimal Incremental Preference Elicitation during Negotiation. In *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina) *(IJCAI'15)*. AAAI Press, 3–9.

[4] Eleni Bardaka, Leila Hajibabai, and Munindar P Singh. 2020. Reimagining ride sharing: Efficient, equitable, sustainable public microtransit. *IEEE Internet Computing* 24, 5 (2020), 38–44.

[5] Craig Boutilier. 2018. Toward user-centric recommender systems. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)*. 2.

[6] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of computational social choice*. Cambridge University Press.

[7] Amit K. Chopra and Munindar P. Singh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 48–53. https://doi.org/10.1145/3278721.3278740

[8] Mehdi Dastani and Vahid Yazdanpanah. 2022. Responsibility of AI Systems. *AI & SOCIETY* (2022), 1–10.

[9] Virginia Dignum. 2017. Responsible autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4698–4704.

[10] Virginia Dignum. 2020. AI is multidisciplinary. *AI Matters* 5, 4 (2020), 18–21.

[11] Maciej Drwal, Enrico Gerding, Sebastian Stein, Keiichiro Hayakawa, and Hironobu Kitaoka. 2017. Adaptive pricing mechanisms for on-demand mobility. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 1017–1025.

[12] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*, Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Eds.). Springer, Berlin, Heidelberg, 1–19.

[13] Dorota Filipczuk, Tim Baarslag, Enrico H. Gerding, and m. c. schraefel. 2022. Automated privacy negotiations with preference uncertainty. *Autonomous Agents and Multi-Agent Systems* 36, 2 (October 2022).

[14] Luciano Floridi. 2018. Soft ethics and the governance of the digital. *Philosophy & Technology* 31, 1 (2018), 1–8.

[15] Luciano Floridi and Josh Cowls. 2022. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design* (2022), 535–545.

[16] Enrico H. Gerding, Valentin Robu, Sebastian Stein, David C. Parkes, Alex Rogers, and Nicholas R. Jennings. 2011. Online Mechanism Design for Electric Vehicle Charging. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2* (Taipei, Taiwan) *(AAMAS '11)*. 811–818.

[17] Omri Gillath, Ting Ai, Michael S. Branicky, Shawn Keshmiri, Robert B. Davison, and Ryan Spaulding. 2021. Attachment and trust in artificial intelligence. *Computers in Human Behavior* 115 (2021), 106607. https://doi.org/10.1016/j.chb.2020.106607

[18] Richard Gomer, m.c. schraefel, and Enrico Gerding. 2014. Consenting agents: semi-autonomous interactions for ubiquitous consent. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct publication*. 653–658.

[19] Fariba Goodarzian, Ata Allah Taleizadeh, Peiman Ghasemi, and Ajith Abraham. 2021. An integrated sustainable medical supply chain network during COVID-19. *Engineering Applications of Artificial Intelligence* 100 (2021), 104188. https://doi.org/10.1016/j.engappai.2021.104188

[20] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. 2022. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences* 12, 3 (2022). https://doi.org/10.3390/app12031353

[21] Tatsuya Iwase, Sebastian Stein, and Enrico H. Gerding. 2021. A Polynomial-time, Truthful, Individually Rational and Budget Balanced Ridesharing Mechanism. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, Zhi-Hua Zhou (Ed.). 268–275. https://doi.org/10.24963/ijcai.2021/38

[22] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. 2017. An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology (TOIT)* 17, 3 (2017), 1–22.

[23] Ilir Kola, Ralvi Isufaj, and Catholijn M Jonker. 2022. Does Personalization Help? Predicting How Social Situations Affect Personal Values. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 157–170.

[24] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[25] Mengya Liu, Vahid Yazdanpanah, Sebastian Stein, and Enrico H. Gerding. 2022. Multiobjective Routing in Sustainable Mobility-On-Demand. In *ATT'22: Workshop Agents in Traffic and Transportation, July 25, 2022, Vienna, Austria: Part of IJCAI 2022*.

[26] Albert Meijer and Martijn Wessels. 2019. Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration* 42, 12 (2019), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

[27] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[28] 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*. 1706–1710.

[29] Pradeep K Murukannaiah and Munindar P Singh. 2020. From machine ethics to internet ethics: Broadening the horizon. *IEEE Internet Computing* 24, 03 (2020), 51–57.

[30] Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA.

[31] Kieron O'Hara. 2019. Data trusts: Ethics, architecture and governance for trustworthy data stewardship. (2019).

[32] Sascha Ossowski. 2012. *Agreement technologies*. Vol. 8. Springer Science & Business Media.

[33] Ben Purvis, Yong Mao, and Darren Robinson. 2019. Three pillars of sustainability: in search of conceptual origins. *Sustainability science* 14 (2019), 681–695.

[34] Sarvapali D. Ramchurn, Trung Dong Huynh, Feng Wu, Yuki Ikuno, Jack Flann, Luc Moreau, Joel Fischer, Wenchao Jiang, Tom Rodden, Edwin Simpson, Steven Reece, Stephen Roberts, and Nicholas R. Jennings. 2016. A disaster response system based on human-agent collectives. *Journal of Artificial Intelligence Research* 57 (2016), 661–708.

[35] Sarvapali D Ramchurn, Sebastian Stein, and Nicholas R Jennings. 2021. Trustworthy human-AI partnerships. *iScience* 24, 8 (2021), 102891.

[36] Sarvapali D Ramchurn, Perukrishnen Vytelingum, Alex Rogers, and Nicholas R Jennings. 2012. Putting the'smarts' into the smart grid: a grand challenge for artificial intelligence. *Commun. ACM* 55, 4 (2012), 86–97.

[37] Emmanouil S. Rigas, Enrico H. Gerding, Sebastian Stein, Sarvapali D. Ramchurn, and Nick Bassiliades. 2022. Mechanism design for efficient offline and online allocation of electric vehicles to charging stations. *Energies* 15, 5 (March 2022).

[38] Edwina L Rissland, Kevin D Ashley, and Ronald Prescott Loui. 2003. AI and Law: A fruitful synergy. *Artificial Intelligence* 150, 1-2 (2003), 1–15.

[39] Stuart Russell. 2022. Provably beneficial artificial intelligence. In *27th International Conference on Intelligent User Interfaces*. 3.

[40] Jordi Sabater and Carles Sierra. 2005. Review on computational trust and reputation models. *Artificial intelligence review* 24, 1 (2005), 33–60.

[41] Amirali Salehi-Abari and Craig Boutilier. 2015. Preference-Oriented Social Networks: Group Recommendation and Inference. In *Proceedings of the 9th ACM Conference on Recommender Systems* (Vienna, Austria) *(RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 35–42. https://doi.org/10.1145/2792838.2800190

[42] Anna Sepliarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. 2018. Preference Elicitation as an Optimization Problem. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 172–180. https://doi.org/10.1145/3240323.3240352

[43] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial Intelligence for Social Good: A Survey. https://doi.org/10.48550/ARXIV.2001.01818

[44] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

[45] Amika M. Singh and Munindar P. Singh. 2023. Wasabi: A Conceptual Model for Trustworthy Artificial Intelligence. *Computer* 56, 2 (2023), 20–28. https://doi.org/10.1109/MC.2022.3212022

[46] Zoi Terzopoulou and Ulle Endriss. 2020. Neutrality and relative acceptability in judgment aggregation. *Social Choice and Welfare* 55, 1 (2020), 25–49.

[47] Sander van Cranenburgh, Shenhao Wang, Akshay Vij, Francisco Pereira, and Joan Walker. 2022. Choice modelling in the age of machine learning - Discussion paper. *Journal of Choice Modelling* 42 (2022), 100340. https://doi.org/10.1016/j.jocm.2021.100340

[48] Max Van Kleek and Kieron O'Hara. 2014. *The Future of Social Is Personal: The Potential of the Personal Data Store*. Springer International Publishing, Cham, 125–158. https://doi.org/10.1007/978-3-319-08681-1_7

[49] Moshe Y Vardi. 2020. Efficiency vs. resilience: what COVID-19 teaches computing. *Commun. ACM* 63, 5 (2020), 3.

[50] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36 (2021), e5. https://doi.org/10.1017/S0269888921000011

[51] Jennifer Williams, Vahid Yazdanpanah, and Stein Sebastian. 2023. Privacy-Preserving Occupancy Estimation. In *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, in–press.

[52] Jennifer Williams, Vahid Yazdanpanah, and Sebastian Stein. 2022. Safe Audio AI Services in Smart Buildings. In *BuildSys'22*. Association for Computing Machinery, 266–269.

[53] Alan FT Winfield and Marina Jirotka. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180085.

[54] Jessica Woodgate and Nirav Ajmeri. 2022. Macro ethics for governing equitable sociotechnical systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1824–1828.

[55] Wei Xu. 2019. Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions* 26, 4 (jun 2019), 42–46. https://doi.org/10.1145/3328485

[56] Vahid Yazdanpanah, Mehdi Dastani, Shaheen Fatima, Nicholas R. Jennings, Devrim Murat Yazan, and W. Henk Zijm. 2020. Task Coordination in Multiagent Systems. In *Proceedings of AAMAS-2020*. 2056–2058.

[57] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Corina Cîrstea, m. c. schraefel, Timothy J. Norman, and Nicholas R. Jennings. 2021. Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements. *IEEE Internet Comput.* 25, 6 (2021), 15–22. https://doi.org/10.1109/MIC.2021.3107334

[58] Vahid Yazdanpanah, Sara Mehryar, Nicholas R. Jennings, Swenja Surminski, Martin J. Siegert, and Jos van Hillegersberg. 2020. Multiagent Climate Change Research. In *Proceedings of AAMAS-2020*. 1726–1731.

[59] Elnaz Shafipour Yourdshahi and Sebastian Stein. 2022. *Electric Vehicle Charging on Long Journeys: Current Challenges and Future Opportunities*. Project Report. University of Southampton. https://doi.org/10.5258/SOTON/PP0006

[60] Efstathios Zavvos, Enrico H. Gerding, Vahid Yazdanpanah, Carsten Maple, Sebastian Stein, and m.c. schraefel. 2022. Privacy and Trust in the Internet of Vehicles. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2022), 10126–10141. https://doi.org/10.1109/TITS.2021.3121125

[61] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. 2021. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *Handbook of Reinforcement Learning and Control* 18 (2021), 321.