Adaptive Incentive Engineering in Citizen-Centric AI

Blue Sky Ideas Track

Behrad Koohy
University of Southampton, UK
bk2g18@soton.ac.uk

Jan Buermann
University of Southampton, UK
j.buermann@soton.ac.uk

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

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

Enrico Gerding
University of Southampton, UK
eg@ecs.soton.ac.uk

Paul Pschierer-Barnfather
Zaptec, UK
paul@zaptec.com

Pamela Briggs
University of Northumbria, UK
p.briggs@northumbria.ac.uk

ABSTRACT

Adaptive incentives are a valuable tool shown to improve the efficiency of complex multiagent systems and could produce win-win situations for all stakeholders. However, their application usage is very limited, partly due to a significant gap between the literature and practice. We argue that overcoming this gap requires addressing four open research challenges. First, the dynamic, volatile and uncertain nature of environments needs to be fully considered. Second, social factors including user acceptance, fairness, ethical considerations and trust have to match end users' expectations and needs. Third, the evaluation of mechanisms and systems has to be robust and focused on real-world outcomes and stakeholder requirements. Finally, all this has to be built on a reliable theoretical foundation. In order to overcome these open challenges in adaptive incentive engineering, tools from the fields of mechanism design and game theory can be used. This will help to achieve the opportunities adaptive incentives can provide to real-world practical environments, producing better AI systems for the benefit of all.

KEYWORDS

Citizen-Centric AI; Incentive Engineering; Mechanism Design; Explainability in AI; AI Ethics and Regulation

ACM Reference Format:
Behrad Koohy, Jan Buermann, Sebastian Stein, Vahid Yazdanpanah, Enrico Gerding, Paul Pschierer-Barnfather, and Pamela Briggs. 2024. Adaptive Incentive Engineering in Citizen-Centric AI: Blue Sky Ideas Track. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 6 pages.

1 INTRODUCTION

In modern society, many engineered systems are modelled and tightly interconnected with user behaviour. The performance of these systems, which range from online e-commerce systems to public infrastructure and services, is dependent on the active participation and conduct of the users that engage with these systems. User behaviour serves as a critical determinant influencing the efficacy, adaptability and effectiveness of these systems. Hence, providing users with incentives to participate in these systems or change their behaviour in a socially beneficial way is an effective way to increase participation and engagement. This can result in systems that have society-wide benefits, such as reduced vehicle emissions and health costs [77].

The related topic of dynamic and personalised pricing of goods for consumers has been extensively researched; in 1999, data-driven techniques have been proposed to alter the price of goods based on their utility [80]. Similarly, differential pricing and Adaptive Incentive Engineering (AIE)[1] have the potential to maximise social welfare [91]. Pervasive computing devices, ranging from smartphones to road vehicles, generate a significant amount of data about their users [45], and have become commonplace in some fields including smart grids, mobility markets and crowdsourcing [72], As such, they are a crucial part of providing the data required for effective and efficient AIE. Nevertheless, differential pricing has often been controversial. Notably, the Clayton Act of 1914 intended to ban it to protect small businesses [91] and a more recent attempt by Amazon backfired [53]. At the same time, Mitra [55] found that AIE is acceptable for consumers in industries such as hospitality, where demand can far outstrip supply. This attitude from consumers has expanded to include other industries where consumers are aware of the limited supply, and where AIE is an effective approach to managing demand [80].

We observe a lack of methods and theories within the AI/MAS community for effectively engineering AIE mechanisms in real-world applications. We emphasise the need for robust foundations,

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[1]We refer to Adaptive Incentive Engineering (AIE) as an approach within the domain of incentive design [63, 66] aimed at influencing the behaviour of AI/human agents towards a specific behaviour. Unlike traditional incentive mechanisms that rely on monetary rewards or pricing [23, 92], adaptive incentive engineering encompasses a broader range of incentives, including non-monetary rewards, recognition, and social benefits. This approach involves the continuous fine-tuning of incentives based on real-time feedback and an understanding of the evolving context. The term "adaptive" highlights the responsiveness and flexibility of the incentive system, acknowledging that environmental conditions, users' preferences, and external factors may change over time.

frameworks, and evaluation metrics to develop AIE mechanisms that are optimally feasible in terms of fairness, resilience, and autonomy. We identify challenges from both academic and non-academic perspectives and map them to a solution concept capable of addressing AIE challenges. Our proposed outcome aims to facilitate the effective, reliable, and trustworthy development and deployment of AI, contributing to the AI community while providing solutions to real-world problems.

2 CONCEPTUAL DYNAMICS OF INCENTIVES

As discussed in [84], establishing citizen-centric AI requires ensuring that AI systems are able to address a diverse range of users' preferences and that end-users can actively contribute to the decision-making process. Furthermore, it is essential to provide insights to regulatory bodies and be aware of changes in regulations and guidelines provided by standardisation institutes[2] as well as user and industry representatives. Establishing such multi-directional relations and keeping involved stakeholders satisfied requires incentives in various forms. Incentives are necessary to maintain collaboration among stakeholders and/or to nudge the practice towards social good [87]. Such incentives can take a monetary form (e.g., waiving taxation or providing subsidies) aiming at financial profitability of socially-beneficial practices or could be non-monetary (e.g., reviews, carbon credits, or green badges) aiming at improving the image and reputation of service providers. This highlights the significance of incentives at a macro-economic level. However, in case of AI-based services, incentives are also relevant on the micro-economic level, inside service provider firms given the nature of the AI systems and how they interact with end-users. That is, in contrast to traditional industries and services that used to deploy (non-intelligent) mechanical and reliably predictive technologies, AI-based services use AI agents with a significant degree of autonomy and serve (human) users with diverse and context-dependent preferences. So, in addition to already known challenges related to nudging in general (e.g., see [11]), nudging and engineering incentives in AI-based services lead to new challenges, as the technology itself uses components with a degree of autonomy and agency. To that end, the topic we invite the AI/MAS community to focus on is:

"engineering incentives dynamically to align the behaviour of AI systems with citizen end-users' values and preferences".

This research explores the dynamics of nudging and incentives in human-AI systems, considering ethical aspects [78], desirable properties, and potential tensions. We analyze key characteristics of the context for effective AIE in citizen-centric AI and emphasise desirable properties of solution concepts, addressing tensions and motivating new research in this direction. We define the effectiveness as the ability to complete a task with the minimal use of resources, and efficiency as the degree of which a system or process achieves the desired outcome.

What citizen-centricity demands and key characteristics of the context that shape the scope of AIE: In citizen-centric AI systems [84], we require incentive mechanisms that effectively capture:

- *Interactive Dynamicity* of the system involves dynamic state changes and interactions among diverse human and AI entities, leading to emergent behavior [95]. This dynamic nature in citizen-centric AI creates a temporally evolving target for defining desirable and socially beneficial behavior. Stakeholders face different forms of uncertainty, with some controllable by stakeholders (e.g., service users or providers) and others inherently uncontrollable (e.g., climate unpredictability).
- *Resource Volatility* in how service providers can perform their services, e.g., what mode and level of transportation providers can supply or what level of energy that energy providers have the capacity to provide (this is partially under the control of the service providers, not end-users) [41, 65].
- *Demand Uncertainty* pertains to ambiguity about the accuracy of information from users, like the authenticity of service requests or expected delivery times. This uncertainty raises concerns about users' truthfulness and the potential for strategic manipulation of information in reporting usage or preferences [37, 48]. The uncertainty tied to strategic behaviour adds complexity to achieving effective AIE [83].
- *Environment Unpredictability* encompasses physical-temporal uncertainties caused by natural situations, illustrated by fluctuations in (e.g., transportation or energy) service demands throughout different periods—daily, weekly and annually. Such variations are attributed to factors beyond the users' control, such as weather conditions. The dynamic and unpredictable nature of the environment introduces challenges in anticipating and adapting to shifts in demands, requiring AIE mechanisms that account for external factors and their impacts on the system.

On desirable properties of an effective AIE mechanism: In our perspective, three key properties constitute a pool of desiderata principles with tensions. AIE is expected to optimise and ensure:

- *Fairness* for human users in accessing incentives, but at the same time *diversity* and customisability for different needs in an inclusive way. How fairness of an AIE mechanism can be evaluated and balanced against its capacity to address diverse group of users is key here [17].
- *Resilience* and fault-tolerance of AIEs (e.g., by ring-fencing some incentive resources) but at the same time their *efficiency* and budget-balancedness, which requires balancing between using as few resources as possible, on one side, and adding redundancy to cover unanticipated scenarios, on the other side [90].
- *Autonomy* of users in choosing types of incentives while preserving *stability*, profitability and sustainability of the service, which relies on the tension between incentive coordination measure necessary for ensuring collective good, on one end, and individuals' freedom of choice and flexibility in opting for preferred outcomes, on the other end [47, 86].

3 THEORETICAL FOUNDATIONS OF AIE

AIE has a broad foundation in (algorithmic) mechanisms design [61] and their wider formal economic treatment going back over 200 years [27]. The economic foundation may explain the prominence of passive-user utility-focused money-based mechanisms, which are the economic standard [68]. Nevertheless, in a time of real-world-focused citizen-centric AI systems, the theoretical foundations must

[2]Developing novel regulatory measures is key, as, currently, guidelines around fair and trustworthy AI remain relatively generic and are not supported by regulatory measures, e.g., see [59].

broaden to reduce the gap between theory and practical applications and strengthen the applied research and developed AI systems.

Expansion - determining the right mechanisms: The first step in broadening AIE is a need to expand the variety of mechanisms applicable in diverse as well as restricted settings. Monetary incentives may be inapplicable due to social limitations or users' aversion [31]. Similarly, random mechanisms have a limited usage due to barriers of entry, influenced by users' difficulty of understanding randomness [94]. Alternatives within algorithmic mechanism design without money, like information control [3] and utility limitation [68], are still under-explored, despite the area's introduction over a decade ago [68]. Considering the variety of possible settings, there is a need to devise mechanisms on a case-by-case basis [79].

Representation - modelling users as the active participants they are: Beside the general broadening of AIE mechanism types, AIE requires modelling users as active participants via capturing related behaviours and attitudes, which are misrepresented by numerous concepts in mechanism design. Starting with the fundamental assumptions of rationality [62], it has been shown in numerous studies that users have different degrees of rationality, behave situation dependent [42], and might even behave irrationally [69]. An example that shows the gap in modelling and reality is incentive-compatibility (IC), which is intended to guarantee that users report their preferences truthfully. However, designing mechanisms to be IC is impotent if users do not understand the concept and attempt to gain a better outcome via misreporting [73]. Such lack of understanding is likely only exaggerated by learning-based approaches. In the worst case, this leads to perverse incentives contrary to the designer's aims, rendering AIE at best ineffective [38, 93].

The lack in user representation compounds accepting the efficiency loss from requirements. For example, even if we require IC, the detriment to efficiency [16] is not universal. While non-truthfulness may produce arbitrarily bad results, possibly even under small manipulations [88], in other instances, manipulation might be hard [6, 32] or its benefits diminish in large markets [40]. Moreover, most arguments rely on the common knowledge of all users' preferences [6], which is an unreasonable assumption in real applications. Alternatively, an intended social outcome of serving all vulnerable users might exceed requiring IC [3].

Both of the previous points highlight a trend to focus on mathematically convenient concepts rather than practically useful ones. Another example of this is envy-freeness whose preeminence is undermined by its diminished empirical relevance due to its abstractness [30]. Even the welcome attempt to overcome IC's issue via obviously strategy-proofness (OSP) [49] still follows a mathematically neat robust-optimisation approach rather than a real-user inspired concept. Moreover, this approach may diminish efficiency even more than IC - so far no approximation guarantee is known [20, 21] - and it is incongruent with other concepts - OSP is incompatible with matching stability [4]. Similarly, the common approach of relaxing concepts, e.g., approximate IC [5, 44, 85] or fairness [8], is mathematically sound while it is unclear if this is more user acceptable and if an acceptable degree of relaxation exists [30].

Environment - modelling the world as it is - uncertain: Further extending AIE, application environments, foremost the multi-faceted presence of uncertainty, have to be captured better [10, 12, 42]. A strong worst-cases focus manifests in numerous approaches optimising worst revenues or worst social welfare, or being more robust to uncertainty than deterministic optimal mechanisms [52, 67]. However, those approaches neglect more efficient average cases [56, 57] and the possible utilisation of uncertainty for better outcomes [12, 16, 71]. Contrarily, if uncertainty is a strong detriment to a systems, approaches to mitigate or share its effects deserve attention, in comparison to always having users cover the risk [81].

Economics - modelling markets and economic decision making: Finally, AIE has to consider markets and companies as they operate. Balancing economic and social dimensions, economic viability analyses have to replace common notions like budget balance (BB). For example, BB only considers one-shot offline settings and neither accounts for online long-term planning [62] nor companies' profits aims [18]. This is very limited in comparison to economic appraisal techniques which play a crucial role in a company's planning and investment strategies. Similarly, market interaction modelling, like the existence of alternative mechanisms that are favourable for users [71], or which attempt to lock users in [22], needs to replace simple participation concepts like individual rationality. Overall, AIE has to capture real-world applications to be convincingly presented in their viability and usefulness for all stakeholders.

Recommendations: In summary, for useful efficient AIE, it is clear that theoretical foundations are essential but also that the focus of exploration should shift towards application-oriented mechanisms:

(1) Real user autonomy, rationality, preference, utility and fairness attitudes must be conceptually reflected and investigated.
(2) Uncertainty has to be fully captured and explored in all facets.
(3) Economical decision making must be conceptually captured.
(4) More real-world alignment and average-case analysis should complement worst-case and robust optimisation.
(5) Simplicity or explainability to overcome accessibility issues.
(6) Evaluation methods must focus along the same axes of what users, companies and other stakeholders want, need and require.

4 MULTIAGENT AIE TECHNIQUES

The field of AIE has witnessed substantial advances, propelled by contributions from the multiagent systems and the wider AI community. Section 3 highlights the importance of the theoretical background to incentive engineering, and can capture elements such as uncertainty, economical decision making and real user autonomy.Integrating theoretical incentive engineering with AI techniques enables the abstraction of a citizen's utility or quality function[3] in scenarios where direct observation is impractical. This proves particularly beneficial in dynamic pricing applications, like charging strategies for electric vehicles [2], online marketplaces [46], and ride-sharing platforms [96], where the convexity of the utility function may be unknown, and analytical solutions are computationally challenging.

These techniques are essential due to the dynamic nature and uncertainty inherent in this problem. In the context of ride-sharing, dynamicity arises from temporal changes (variations in demand and resource availability throughout the day), spatial variability

[3]We define these functions as ones which capture stakeholders' preferences and priorities, reflecting the inherent trade-offs and considerations in the decision-making process. Preferences may encompass a range of factors, such as monetary gains, user satisfaction, system efficiency, or other relevant metrics.

(resource availability and demand likelihood), traffic conditions, and user preferences. These multiple sources of dynamicity make analytical solutions infeasible or computationally challenging to achieve. Multiagent learning techniques such as multiagent RL can be used to equip multiagent systems with distributed intelligence that can capture the required understanding and navigate a practical scenario to achieve acceptable outcomes for all stakeholders [50].

Adaptive learning techniques can be used to continually learn and adjust AIE policy and behaviour in real time, tracking and responding to changes in the equilibrium of a highly dynamic system [19], particularly in contexts where the provision of incentives can lead to perturbations within the environment itself (e.g., a poor pricing strategy for ride-sharing applications can lead to significantly increased fares during events such as natural disasters, extreme weather and/or public emergencies). Furthermore, leveraging techniques from multiagent AI, such as transfer learning [9] and Bayesian inference [14], can effectively address uncertainties in the incentive engineering problem, distinct from its dynamic nature. This uncertainty may stem from environmental factors like supply chain disruptions and short-term shifts in consumer behavior. Additionally, uncertainties may arise from user behavior, where a portion of the population may not engage with proposed adaptive incentive systems, provide inaccurate or incomplete information, or behave irrationally.

The solution to this outcome is multifaceted; the adaptive agent is constrained to the provided action space based on the available information. It is the responsibility of the model creators to ensure that the underlying incentive mechanism aligns with goals from mechanism design [61]. In designing the mechanism for applying multiagent learning to this problem, an important consideration involves interactions with competing agents. Research by Kastius et al.[39] reveals that RL agents in oligopolies can unintentionally collude even without direct communication. This finding has significant implications for the real-world implementation of adaptive incentive pricing. Instances of collusion or anti-consumer behaviour can pose challenges to the citizen-centric nature of adaptive incentive engineering, potentially eroding trust and confidence [84].

5 ETHICS AND TRUST IN AIE

While incentivising users toward socially optimal solutions, such as promoting environmentally friendly practices, shows promise, it is crucial to consider the potential impact on user trust. Decisions like setting room temperature, where preferences of different age groups may clash or shared rides with varying prices for different users may raise questions about the rationale behind such differentials, potentially influencing users' trust and adoption of these services. Although the AI/MAS community has developed models of trust and ethicality [13, 60, 70], it is necessary to make such models dynamic to adapt to the incentives these technologies use.

Balancing Fairness and Positive Discrimination: The first challenge is balancing fairness and positive discrimination, as seen in shared rides with varying prices and qualities. Allocating resources proportionally based on needs or contributions may conflict with socially optimal solutions. For example, more prosperous neighbourhoods may demand and afford superior transportation services, potentially leading to unintentional favouritism. These dynamics could also lead to user prejudice and the emergence of discrimination in AI-assisted markets, e.g., the possibility for discrimination in ride-sharing [1, 58]. Tools are needed to evaluate and optimise various objectives and aspects of fairness [51], provide online incentive mechanisms [98], and dynamically balance proportional fairness metrics and equity, ensuring fair allocation while accommodating diverse preferences [7].

Adaptive Incentives and Dynamic Trust Evaluation: The second challenge is around the transparency of changing prices or other types of incentives for AI-assisted services and the subsequent impact on user trust. To address this, it is crucial to shift focus from purely history-oriented perspectives on trust modelling that model trust in an AI agent based how they have performed [34] by integration of the current state of the system, consequently evaluating what agents can actually (in the sense of [29]) deliver in prospect [76]. Moreover, understanding the notion of legitimacy, weighing different preferences, and introducing fluidity in votes/preferences, as seen in liquid democracy [25], offer insights into dynamic trust evaluation. Exploring how interventions can improve social choice methods aligns with the transparency of governance systems, ensuring a more comprehensive understanding of the evolving dynamics of trust in relation to adaptive incentives.

Ethics Customisation: The final challenge involves navigating the diverse ethical perspectives of users [43], and their understanding of privacy, in AI-supported services. Instead of imposing a unified ethical stance, allowing users to customise their preferences within socially acceptable bounds is proposed. This customisation is in particular crucial in cases of flexible autonomy [35] where multiple human/AI agents share control, and different agents may have distinct ethical considerations. For instance, one may set preferences to prioritise safety of the vehicle (an those aboard), but others may be keen to prioritise a maximising the overall safety including safety of pedestrians. Dynamic governance models and fluid democracy approaches [28] present potential solutions to address the challenges of customising ethics. How AI-assisted services should decide in such dilemmatic situations requires novel ethics evaluation services [82] and incorporates various aspects such as dynamic evaluation of trust and responsibility in human-AI systems [36, 97]. This requires methods to effectively evaluate legality of decisions made by, or in collaboration with, AI systems [64] and interdisciplinary approaches to develop legal principles for governing AI.

6 CONCLUSIONS

Our vision is a paradigm of adaptive incentive engineering that is adaptive to the dynamism and uncertainty of the world, respects users as individuals with varying needs and attitudes, and aligns system operators', stakeholders' and end-users' needs and requirements. Adopting this paradigm will help to create beneficial AI systems that support sustainable economic growth while addressing concerns around fairness, trust, explainabilty, bias and manipulation, also mentioned in the Bletchley Declaration [15]. Adapting our research agenda will complement ongoing efforts on responsible AI and pursue incentive engineering along its principles which are generally seen as essential by academia [74, 84, 89], governments [15, 33], NGOs [24, 75] and industry [26, 54, 75].

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 the AutoTrust platform grant (EP/R029563/1).

REFERENCES

[1] Olga Abramova. 2022. No matter what the name, we're all the same? Examining ethnic online discrimination in ridesharing marketplaces. *Electronic Markets* 32, 3 (2022), 1419–1446.

[2] Adil Amin, Wajahat Ullah Khan Tareen, Muhammad Usman, Haider Ali, Inam Bari, Ben Horan, Saad Mekhilef, Muhammad Asif, Saeed Ahmed, and Anzar Mahmood. 2020. A review of optimal charging strategy for electric vehicles under dynamic pricing schemes in the distribution charging network. *Sustainability* 12, 23 (2020), 10160.

[3] Jerry Anunrojwong, Krishnamurthy Iyer, and Vahideh Manshadi. 2020. Information Design for Congested Social Services: Optimal Need-Based Persuasion. *CoRR* abs/2005.0 (may 2020). arXiv:2005.07253

[4] Itai Ashlagi and Yannai A Gonczarowski. 2018. Stable matching mechanisms are not obviously strategy-proof. *Journal of Economic Theory* 177 (sep 2018), 405–425. https://doi.org/10.1016/j.jet.2018.07.001

[5] Eduardo M Azevedo and Eric Budish. 2018. Strategy-proofness in the Large. *The Review of Economic Studies* (aug 2018). https://doi.org/10.1093/restud/rdy042

[6] Haris Aziz, Hans Georg Seedig, Jana Karina Von Wedel, and Jana Karina Von Wedel. 2015. On the Susceptibility of the Deferred Acceptance Algorithm. In *Proceedings of the International Joint Conference on AAMAS*, Vol. 2. 939 – 947.

[7] 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.

[8] Siddharth Barman, Ganesh Ghalme, Shweta Jain, Pooja Kulkarni, and Shivika Narang. 2019. Fair Division of Indivisible Goods Among Strategic Agents. In *Proceedings of the International Joint Conference on AAMAS*. IFAAMAS, Richland, SC, 1811–1813. arXiv:arXiv:1901.09427v1

[9] Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. 2022. Meta dynamic pricing: Transfer learning across experiments. *Management Science* 68, 3 (2022), 1865–1881.

[10] Sweta Bhattacharya, Rajeswari Chengoden, Gautam Srivastava, Mamoun Alazab, Abdul Rehman Javed, Nancy Victor, Praveen Kumar Reddy Maddikunta, and Thippa Reddy Gadekallu. 2022. Incentive Mechanisms for Smart Grid: State of the Art, Challenges, Open Issues, Future Directions. *Big Data and Cognitive Computing* 6, 2 (apr 2022), 47. https://doi.org/10.3390/bdcc6020047

[11] Luc Bovens. 2009. The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology*. Springer, 207–219.

[12] Jan Buermann, Enrico H. Gerding, and Baharak Rastegar. 2020. Fair Allocation of Resources with Uncertain Availability. In *Proc. of the 19th International Conference on AAMAS*. 9 pages.

[13] Cristiano Castelfranchi and Rino Falcone. 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*. IEEE, 72–79.

[14] Xi Chen, Jianjun Gao, Dongdong Ge, and Zizhuo Wang. 2022. Bayesian dynamic learning and pricing with strategic customers. *Production and Operations Management* 31, 8 (2022), 3125–3142.

[15] Countries Attending the AI Safety Summit. 2023. *The Bletchley Declaration*. Technical Report. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[16] Luciano De Castro and Nicholas C. Yannelis. 2018. Uncertainty, efficiency and incentive compatibility: Ambiguity solves the conflict between efficiency and incentive compatibility. *Journal of Economic Theory* 177 (sep 2018), 678–707. https://doi.org/10.1016/j.jet.2018.02.008

[17] Ghislain Herman Demeze-Jouatsa, Roland Pongou, and Jean-Baptiste Tondji. 2022. Justice, inclusion, and incentives. *Available at SSRN 4191231* (2022).

[18] Stylianos Despotakis, R. Ravi, and Amin Sayedi. 2021. First-Price Auctions in Online Display Advertising. *Journal of Marketing Research* 58, 5 (oct 2021), 888–907. https://doi.org/10.1177/0022243721030201/FORMAT/EPUB

[19] Maciej Drwal, Enrico Gerding, Sebastian Stein, Keiichiro Hayakawa, and Hironobu Kitaoka. 2017. Adaptive pricing mechanisms for on-demand mobility. (2017).

[20] Diodato Ferraioli and Carmine Ventre. 2022. Obvious Strategyproofness, Bounded Rationality and Approximation. *Theory of Computing Systems* 66, 3 (jun 2022), 696–720. https://doi.org/10.1007/s00224-022-10071-2

[21] Diodato Ferraioli, Carmine Ventre, Itai Ashlagi, and Yannai A. Gonczarowski. 2017. Obvious Strategyproofness Needs Monitoring for Good Approximations. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, Vol. 177. 516 – 522.

[22] Financial Conduct Authority (FCA). 2018. *Pricing practices in the retail general insurance sector: Household insurance*. Technical Report. 1–27 pages.

[23] MA Gibney, Nicholas R Jennings, NJ Vriend, and José-Marie Griffiths. 1999. Market-based call routing in telecommunications networks using adaptive pricing and real bidding. In *Intelligent Agents for Telecommunication Applications: Third International Workshop, IATA'99*. Springer, 46–61.

[24] Global Partnership on Artificial Intelligence. 2023. *Responsible AI Working Group Report*. Technical Report. https://www.gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-november-2021.pdf

[25] Paul Gölz, Anson Kahng, Simon Mackenzie, and Ariel D Procaccia. 2021. The fluid mechanics of liquid democracy. *ACM Transactions on Economics and Computation* 9, 4 (2021), 1–39.

[26] Google AI. 2022. Google Responsible AI Practices. https://ai.google/responsibility/responsible-ai-practices/

[27] Theodore Groves and John Ledyard. 1988. *Incentive Compatibility: Ten Years Later*. University of Minnesota Press. 48–111 pages.

[28] Daniel Halpern, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, and Manon Revel. 2023. In Defense of Liquid Democracy. In *Proceedings of the 24th ACM Conference on Economics and Computation* (London, United Kingdom) *(EC '23)*. ACM, New York, NY, USA, 852. https://doi.org/10.1145/3580507.3597817

[29] Joseph Y Halpern. 2016. *Actual causality*. MIT Press.

[30] Dorothea K. Herreiner and Clemens D. Puppe. 2009. Envy Freeness in Experimental Fair Division Problems. *Theory and Decision* 67, 1 (jul 2009), 65–100. https://doi.org/10.1007/s11238-007-9069-8

[31] Katelin Hoskins, Connie M. Ulrich, Julianna Shinnick, and Alison M. Buttenheim. 2019. Acceptability of financial incentives for health-related behavior change: An updated systematic review. *Preventive Medicine* 126 (sep 2019), 105762. https://doi.org/10.1016/j.ypmed.2019.105762

[32] Hadi Hosseini, Fatima Umar, and Rohit Vaish. 2022. Two for One & One for All: Two-Sided Manipulation in Matching Markets. In *IJCAI International Joint Conference on Artificial Intelligence*. 321 – 327.

[33] House of Commons Science Innovation and Technology Committee. 2023. *The governance of artificial intelligence: interim report*. Technical Report.

[34] Trung Dong Huynh, Nicholas R Jennings, and Nigel R Shadbolt. 2006. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 13 (2006), 119–154.

[35] Nicholas R Jennings, Luc Moreau, David Nicholson, Sarvapali Ramchurn, Stephen Roberts, Tom Rodden, and Alex Rogers. 2014. Human-agent collectives. *Commun. ACM* 57, 12 (2014), 80–88.

[36] Susanne Kalenka and Nicholas R Jennings. 1999. Socially responsible decision making by autonomous agents. In *Cognition, Agency and Rationality: Proceedings of the Fifth International Colloquium on Cognitive Science*. Springer, 135–149.

[37] Ece Kamar and Eric Horvitz. 2012. Incentives for truthful reporting in crowdsourcing.. In *AAMAS*, Vol. 12. 1329–1330.

[38] Karthik Kannan, Rajib L. Saha, and Warut Khern-am nuai. 2022. Identifying Perverse Incentives in Buyer Profiling on Online Trading Platforms. *Information Systems Research* 33, 2 (jun 2022), 464–475. https://doi.org/10.1287/isre.2021.1077

[39] Alexander Kastius and Rainer Schlosser. 2021. Dynamic pricing under competition using reinforcement learning. *Journal of Revenue and Pricing Management* (2021), 1–14.

[40] John Kennes, Daniel Monte, and Norovsambuu Tumennasan. 2019. Strategic Performance of Deferred Acceptance in Dynamic Matching Problems. *American Economic Journal: Microeconomics* 11, 2 (may 2019), 55–97. https://doi.org/10.1257/mic.20170077

[41] Shinya Kikuchi. 2005. Study of transportation and uncertainty. In *Applied Research in Uncertainty Modeling and Analysis*. Springer, 303–319.

[42] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2023. The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization. *Management Science* (nov 2023). https://doi.org/10.1287/mnsc.2022.03683

[43] Nadin Kokciyan, Biplav Srivastava, Michael N Huhns, and Munindar P Singh. 2021. Sociotechnical perspectives on AI ethics and accountability. *IEEE Internet Computing* 25, 6 (2021), 5–6.

[44] Anshul Kothari, David C Parkes, and Subhash Suri. 2005. Approximately-strategyproof and tractable multiunit auctions. *Decision Support Systems* 39, 1 (mar 2005), 105–121. https://doi.org/10.1016/j.dss.2004.08.009

[45] Abhishek Kumar, Tristan Braud, Sasu Tarkoma, and Pan Hui. 2020. Trustworthy AI in the Age of Pervasive Computing and Big Data. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 1–6. https://doi.org/10.1109/PerComWorkshops48775.2020.9156127

[46] Linchi Kwok and Karen L Xie. 2019. Pricing strategies on Airbnb: Are multi-unit hosts revenue pros? *International Journal of Hospitality Management* 82 (2019), 252–259.

[47] Minae Kwon, John Agapiou, Edgar Duéñez-Guzmán, Romuald Elie, Georgios Piliouras, Kalesha Bullard, and Ian Gemp. 2023. Auto-aligning multiagent incentives with global objectives. In *ALA Workshop, AAMAS*. 1–9.

[48] Maria Kyropoulou and Carmine Ventre. 2019. Obviously strategyproof mechanisms without money for scheduling. In *Proceedings of the International Joint Conference on AAMAS*, Vol. 3. Association for Computing Machinery (ACM), 1574–1581.

[49] Shengwu Li. 2017. Obviously Strategy-Proof Mechanisms. *American Economic Review* 107, 11 (nov 2017), 3257–3287. https://doi.org/10.1257/aer.20160425

[50] Tao Li, Yuhan Zhao, and Quanyan Zhu. 2022. The role of information structures in game-theoretic multi-agent learning. *Annual Reviews in Control* 53 (2022), 296–314.

[51] Mengya Liu, Vahid Yazdanpanah, Sebastian Stein, and Enrico Gerding. 2023. Sustainability-oriented route generation for ridesharing services. *Computer Science and Information Systems* 00 (2023), 53–77.

[52] Giuseppe Lopomo, Luca Rigotti, and Chris Shannon. 2021. Uncertainty in Mechanism Design. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3774581

[53] Jennifer Lyn Cox. 2001. Can differential prices be fair? *Journal of Product & Brand Management* 10, 5 (sep 2001), 264–275. https://doi.org/10.1108/10610420110401829

[54] Meta. 2021. Facebook's five pillars of Responsible AI. https://ai.meta.com/blog/facebooks-five-pillars-of-responsible-ai/

[55] Subrata Kumar Mitra. 2020. An analysis of asymmetry in dynamic pricing of hospitality industry. *International Journal of Hospitality Management* 89 (2020), 102406.

[56] Barnabé Monnot, Francisco Benita, and Georgios Piliouras. 2017. How bad is selfish routing in practice? *CoRR* abs/1703.0 (mar 2017), 93–102. arXiv:1703.01599

[57] Barnabé Monnot, Francisco Benita, and Georgios Piliouras. 2022. Routing Games in the Wild: Efficiency, Equilibration, Regret, and a Price of Anarchy Bound via Long Division. *ACM Transactions on Economics and Computation* 10, 1 (mar 2022), 1–26. https://doi.org/10.1145/3512747

[58] Joanna Moody, Scott Middleton, and Jinhua Zhao. 2019. Rider-to-rider discriminatory attitudes and ridesharing behavior. *Transportation Research Part F: Traffic Psychology and Behaviour* 62 (2019), 258–273.

[59] Luke Munn. 2023. The uselessness of AI ethics. *AI and Ethics* 3, 3 (2023), 869–877.

[60] Pradeep K Murukannaiah, Nirav Ajmeri, Catholijn M Jonker, and Munindar P Singh. 2020. New foundations of ethical multiagent systems. In *Proceedings of the 19th Conference on AAMAS*.

[61] N. Nisan. 2007. Introduction to Mechanism Design (for Computer Scientists). In *Algorithmic Game Theory*, Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani (Eds.). Cambridge University Press, Cambridge, Chapter 9, 209–242. https://doi.org/10.1017/CBO9780511800481

[62] Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511800481

[63] Dmitry A Novikov. 2016. Incentive mechanisms for multi-agent organizational systems. *New Frontiers in Information and Production Systems Modelling and Analysis: Incentive Mechanisms, Competence Management, Knowledge-based Production* (2016), 35–57.

[64] Daria Onitiu, Vahid Yazdanpanah, Age Chapman, Enrico Gerding, Stuart E Middleton, and Jennifer Williams. 2023. On the legal aspects of responsible AI: adaptive change, human oversight, and societal outcomes. In *International Conference on AI for People: Democratizing AI.*

[65] Leonardo Paoli, Richard C Lupton, and Jonathan M Cullen. 2018. Useful energy balance for the UK: An uncertainty analysis. *Applied Energy* 228 (2018), 176–188.

[66] David C Parkes, Ruggiero Cavallo, Florin Constantin, and Satinder Singh. 2010. Dynamic incentive mechanisms. *Ai Magazine* 31, 4 (2010), 79–94.

[67] Georgios Piliouras, Evdokia Nikolova, and Jeff S. Shamma. 2016. Risk Sensitivity of Price of Anarchy Under Uncertainty. *ACM Trans. Econ. Comput.* 5, 1 (2016), 1–27. https://doi.org/10.1145/2930956

[68] Ariel D Procaccia and Moshe Tennenholtz. 2009. Approximate Mechanism Design Without Money. In *Proceedings of the 10th ACM Conference on Electronic Commerce*. ACM, New York, NY, USA, 177–186. https://doi.org/10.1145/1566374.1566401

[69] Benliu Qiu, Yuejiang Lit, Yan Chen, and H. Vicky Zhao. 2019. Controlling Information Diffusion with Irrational Users. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 482–485. https://doi.org/10.1109/APSIPAASC47483.2019.9023355

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

[71] Lillian J Ratliff, Roy Dong, Shreyas Sekar, and Tanner Fiez. 2019. A Perspective on Incentive Design: Challenges and Opportunities. *Annual Review of Control, Robotics, and Autonomous Systems* 2, 1 (may 2019), 305–338. https://doi.org/10.1146/annurev-control-053018-023634

[72] Lillian J Ratliff, Roy Dong, Shreyas Sekar, and Tanner Fiez. 2019. A perspective on incentive design: Challenges and opportunities. *Annual Review of Control, Robotics, and Autonomous Systems* 2 (2019), 305–338.

[73] Alex Rees-Jones and Samuel Skowronek. 2018. An experimental investigation of preference misrepresentation in the residency match. *Proceedings of the National Academy of Sciences* 115, 45 (nov 2018), 11471–11476. https://doi.org/10.1073/pnas.1803212115

[74] Responsible AI UK. 2023. Mission. https://www.rai.ac.uk/mission

[75] Becca Ricks and Mark Surman. 2020. *Creating Trustworthy AI - a Mozilla white paper on challenges and opportunities in the AI era.* Technical Report. Mozilla Foundation.

[76] Asieh Salehi Fathabadi and Vahid Yazdanpanah. 2023. Trust modelling and verification using Event-B. In *Proceedings of the Fifth International Workshop on Formal Methods for Autonomous Systems.*

[77] G Santos, L Rojey, DM Newbery, et al. 2000. *The Environmental Benefits from Road Pricing.* Technical Report. Faculty of Economics, University of Cambridge.

[78] Andreas T Schmidt and Bart Engelen. 2020. The ethics of nudging: An overview. *Philosophy compass* 15, 4 (2020), e12658.

[79] Andreas T. Schmidt and Bart Engelen. 2020. The ethics of nudging: An overview. *Philosophy Compass* 15, 4 (apr 2020). https://doi.org/10.1111/phc3.12658

[80] Peter Seele, Claus Dierksmeier, Reto Hofstetter, and Mario D Schultz. 2021. Mapping the ethicality of algorithmic pricing: A review of dynamic and personalized pricing. *Journal of Business Ethics* 170 (2021), 697–719.

[81] Agus Setiawan, Sugiarto Sugiarto, Grace Shinta S. Ugut, and Edison Hulu. 2021. Fair pricing: A framework towards sustainable life insurance products. *Accounting* 7, 1 (jan 2021), 1–12. https://doi.org/10.5267/j.ac.2020.10.020

[82] Amika M Singh and Munindar P Singh. 2023. Norm deviation in multiagent systems: A foundation for responsible autonomy. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 32.

[83] Faith Jordan Srour and Neil Yorke-Smith. 2018. On Collusion and Coercion: Agent Interconnectedness and In-Group Behaviour.. In *AAMAS*. 1622–1630.

[84] Sebastian Stein and Vahid Yazdanpanah. 2023. Citizen-Centric Multiagent Systems. In *Proceedings of the 2023 International Conference on AAMAS*. 1802–1807.

[85] Xin Sui and Craig Boutilier. 2015. Approximately Strategy-proof Mechanisms for (Constrained) Facility Location. In *Proceedings of the International Joint Conference on AAMAS*. 605 – 613.

[86] K Suzanne Barber, Anuj Goel, and Cheryl E Martin. 2000. Dynamic adaptive autonomy in multi-agent systems. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 2 (2000), 129–147.

[87] Richard H Thaler and Cass R Sunstein. 2021. *Nudge: The final edition.* Yale University Press.

[88] Rohit Vaish and Dinesh Garg. 2017. Manipulating Gale-Shapley Algorithm: Preserving Stability and Remaining Inconspicuous. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, California, 437–443. https://doi.org/10.24963/ijcai.2017/62

[89] Shannon Vallor, Joanna Al-Qaddoumi, Stuart Anderson, Vaishak Belle, Michael Fisher, Bhargavi Ganesh, Ibrahim Habli, Louise Hatherall, Richard Hawkins, Marina Jirotka, Dilara Keküllüoğlu, Nadin Kokciyan, Lars Kunze, John McDermid, Phillip Morgan, Sarah Moth-Lund Christensen, Paul Noordhof, Zoe Porter, Michael Rovatsos, Nayha Sethi, Jack Stilgoe, Carolyn Ten Holter, Tillmann Vierkant, and Robin Williams. 2023. *Edinburgh Declaration on Responsibility for Responsible AI.* Technical Report. https://medium.com/@svallor{_}10030/edinburgh-declaration-on-responsibility-for-responsible-ai-1a98ed2e328b

[90] Moshe Y Vardi. 2022. Efficiency vs. resilience: Lessons from COVID-19. *Perspectives on digital humanism* (2022), 285–289.

[91] Hal R. Varian. 1989. Chapter 10 Price discrimination. In *Handbook of Industrial Organization*. Vol. 1. Elsevier, 597–654. https://doi.org/10.1016/S1573-448X(89)01013-7

[92] Thomas Voice, Perukrishnen Vytelingum, Sarvapali Ramchurn, Alex Rogers, and Nicholas Jennings. 2011. Decentralised control of micro-storage in the smart grid. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 25. 1421–1427.

[93] Gro Holst Volden. 2019. Public funding, perverse incentives, and counterproductive outcomes. *International Journal of Managing Projects in Business* 12, 2 (jun 2019), 466–486. https://doi.org/10.1108/IJMPB-12-2017-0164

[94] Joseph Jay Williams and Thomas L Griffiths. 2013. Why are people bad at detecting randomness? A statistical argument. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39, 5 (2013), 1473–1490. https://doi.org/10.1037/a0032397

[95] Paul Pao-Yen Wu, Clinton Fookes, Jegar Pitchforth, and Kerrie Mengersen. 2015. A framework for model integration and holistic modelling of socio-technical systems. *Decision Support Systems* 71 (2015), 14–27.

[96] Chiwei Yan, Helin Zhu, Nikita Korolko, and Dawn Woodard. 2020. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)* 67, 8 (2020), 705–724.

[97] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, Timothy J. Norman, and Sarvapali D. Ramchurn. 2023. Reasoning about responsibility in autonomous systems: challenges and opportunities. *AI Soc.* 38, 4 (2023), 1453–1464. https://doi.org/10.1007/S00146-022-01607-8

[98] Hanrui Zhang and Vincent Conitzer. 2021. Automated dynamic mechanism design. *Advances in Neural Information Processing Systems* 34 (2021), 27785–27797.