AI and Ethics (2023) 3:1447–1454
https://doi.org/10.1007/s43681-022-00253-6

ORIGINAL RESEARCH

Engineering a social contract: Rawlsian distributive justice through algorithmic game theory and artificial intelligence

Hutan Ashrafian[1]

Received: 11 June 2022 / Accepted: 8 December 2022 / Published online: 22 December 2022

Abstract

The potential for artificial intelligence algorithms and game theory concepts to offer prescriptive and decision-making capability for humankind is increasingly recognized. This derives from the increasing availability of granular, multivariable, well-curated data offering analytical insights for necessarily complex human behaviors and activities. Of the multitude of situations that this decision-making aptitude presents, the application to governmental policy offers a commanding case. This would allow decisions to be made for the benefit of societies and citizens based on rigorous objective information devoid of the traditional approach of choosing policies and societal values based on the opinion of a handful of selected representatives who may be exposed to a lack of comprehensive data analysis capacity and subject to personal biases. There would need to be a critical requirement of wider socially responsible data practices here, beyond those of technical considerations and the incorporation of wider societal fairness approaches. Amongst the schools of political thought particularly acquiescent to the application by this approach would be the egalitarian approach of John Rawls. Here an Original Position's pre-determination tool of Veil of Ignorance and ensuing Difference Principal presents a method of distributive justice that can be clearly mathematically defined in economics theory through Wald's Maximin principle. This offers an opportunity to apply algorithmic game theory and artificial intelligence computational approaches to implement Rawlsian distributive justice that are presented and discussed. The outputs from the algorithmic acquaintance of Rawlsian egalitarianism with applicable state data, protected with appropriate privacy, security, legal, ethical and social governance could in turn lead to automated direct governmental choices and an objective Social Contract for citizens of digitally literate nations.

Keywords Artificial intelligence · Distributive justice · Fairness · Politics · Decision-making · Policy · Rawlsian · 'John Rawls' · Society · Game theory · Algorithm · Algorithmic game theory · Government · Bias · Algorithmic bias · Government · Social contract

1 Exemplum Moralem (a speculative and fictional case example)

Aipotue, geographically part of the larger Danlsi islands group in the Northern Ocean, is among the least populated countries on the planet with its population living on around 300 islands. It had become independent after it had decided against becoming part of Greater Danlsi (the largest and most dominant of the island group) due to cultural and linguistic differences. A widely spread-out nation with a history of socioeconomic benefits and political elitism resting largely with the central islanders, it had been one of the earliest adopters of national digitization to facilitate communication, social and healthcare and financial transactions. The new President who was a champion of technocracy had won his democratic campaign on a promise of a fair rule through the real-world enactment of a new constitution based on a platform of egalitarianism in the absence of expected human biases. After a twelve-month instigation of an artificial intelligent and game theory based system acting through a 'Veil of Ignorance' (that employed the best bias mitigating algorithms available) utilizing all available national data sets, a social contract was developed, new laws written, and a new social order introduced. Politicians on neighboring islands looked on, wondering which other nation state would be the first to follow.

✉ Hutan Ashrafian
hutan@ic.ac.uk

[1] Imperial College London, 10th Floor, Queen Elizabeth the Queen Mother (QEQM) Building, South Wharf Road, London W2 1NY, UK

2 Introduction

Artificial intelligence and game theory have been increasingly applied to both the theoretical and practical implementation of law, health, society, and politics so that it can also be applied to the foundations of these through justice. The forthcoming opportunity to enhance on current machine learning approaches and eventually, one day to achieve Artificial General Intelligence with comparative human sentience, cognition and perception [3, 4] offers current and future AI tools to develop solutions for a social contract to realize its progression from Immanuel Kant's position with the presumption of limits on the state to the work of John Rawls' contractarian theory of justice. It may however be possible to achieve many social contract solutions though contemporary artificial intelligence and game theory approaches.

Theories of justice have been considered and evolved over time and just as artificial intelligence has floundered and flourished and morphed over the past fifty years (surviving two AI winters before sparking the current 'fourth industrial revolution'), in a comparable way theories of justice have also waxed and waned. Glaucon's postulate in Plato's Republic (375BC) and the Lokottaravāda text of Mahāvastu (dated between the second century BCE to fourth century CE) offer some of the earliest seeds of a social contract, though after several 'winters' the more recent ideas of Hobbes, Locke, Rousseau (and various interpretations of utilitarianism and justice through the lens of the common good) to Rawls (where justice was presented as a fairness) and those who follow have now kept the idea of a social contract for justice as a core element of societal philosophical reflection. The Rawlsian idea [20] which is well recognized and prized in its concept of an original position and 'veil of ignorance' and his Difference Principle are distinctive as they can be captured in a simple thought experiment that can be algorithmically characterized, interpreted, and actioned. This conceptual approach currently offers a distinctive opportunity (more so than other more diffuse theories of justice) to allow an artificially generated conscious agent to participate in Rawlsian-based game theory process to solve the thought experiment to achieve the function of justice for all individuals in a society. The aim of this manuscript is to characterize the possible methodological solutions to Rawls' thought experiment with barriers and solutions to achieving justice and fairness through contemporary game theory and artificial intelligence approaches.

3 Rawls-justice as fairness

Rawls highlighted that in a utilitarian system, government has a responsibility to structure society with a functional aim of maximizing production and optimize the distribution of welfare so that if inequality arises, it would be considered an acceptable sacrifice 'for the greater good' [20]. He offered a governmental solution that would consider every individual need so that there would not need to be an inequality and every member of society would have access to basic needs such as essential goods, access to education and social mobility. Achieving this would then offer a just society, and governments would attain this by restructuring their distribution of goods through a means of redistributive justice.

To generate a solution for this, he offered a hierarchy of needs that was necessary for every individual in society: (i) freedom, (ii) equal opportunity (in terms of all/any advantages in terms of resources or privileges) and (iii) the difference principle. The latter stipulated that given these first two principles are met, then an unequal society is acceptable if the system functions to benefit the least privileged.

Solving the needs from this hierarchy answered practical questions such as how to create a guarantee for basic rights, rights to goods, approaches to taxation and welfare. By doing so, he moved away from utilitarianism to interpret justice as fairness.

Rawls' solution was his thought experiment where individuals tasked with setting governmental rules for distributive justice would start in an 'original position' with a 'veil of ignorance' (not knowing or being able to guess or predict what position they would hold in this society). With this imaginary veil, each person decides on the basic goods and privileges to which any member of society is entitled—not knowing whether they will be recipients where in the societal hierarchy they will be or what goods or aid they will receive. As a result, as everyone in the original position is coming from a place of self-interest without knowing what eventual position they will hold in society, their decisions about entitlement to basic goods and privileges be just (as they might have to exist with any status in that society). Here, even if there were some inequalities in society the veil of ignorance would render them acceptable, as they would be devoid of biases on inherent social standing, wealth, or privilege. His argument characterizes the situation that if an individual in the original position found themselves at the bottom of a social hierarchy, then that would acceptable because the decisions that led to the rules of redistribution were made without knowing who was going to be at the top or bottom of the hierarchy independent to the political stance or economic status of the state.

Importantly, Rawls introduced the concept of Wald's Maximin [21] in his Difference Principle, suggesting that deliberations for choosing a society could also consider inequalities that would offer a tangible real-world solution. Here he posited that should any scenario for an inequality to exist, the genesis of that inequality could only be introduced if the worst off also benefited from that system. As a result, there would be a weighted function (W) of resource allocation across the population that would ensure that if there was any difference in resource allocation in a society with so-called 'winners' (with the highest growth and share of the resources) the worst-off (min) segment of the population in terms of utility (u) would also be benefit to as much as possible (the maximum level).

$$W\left(u_1, u_2, \ldots, u_n\right) = \min\{u_1, u_2, \ldots, u_n\}.$$

The Wald's Maximin was the source of significant debate by Rawls and economic Nobel laureate John Harsanyi, questioning whether this approach can act as a route to achieving morality [10], it is nonetheless accepted as a representative resource allocation instrument of Rawl's difference principle. Consequently, whilst there is a myriad of egalitarian philosophies in existence that continue to promulgate, Rawls' system remains the most tangibly algorithmic to allow an A system to offer a Rawlsian simulation of the veil of ignorance and the difference principle (if adequate data was offered for decisions). Thus, whilst arguments over what measures of fairness exist, what are acceptable disparities if at all? Should fairness be an equal probability function for all events, how does the worst off get treated and what exact risk benefits are acceptable in any society, one interpretation of Rawls can bypass many of these issues by an iterative selection of n individuals to select a just society, knowing all the variables of current society and selecting appropriate life journey that are acceptable behind a veil. By doing so this interpretation of Rawls offers a solution to be made synthetically by an artificially intelligent agent.

The application of artificial intelligent agents to generate consequential decisions of justice for populations requires the application of AI (such as machine learning-ML) models (in the current era). Programming an algorithm to offer the 'Original Position' and generate a 'veil of ignorance' and enact the Difference Principle, however, is conceptually feasible and plausible. Whilst a futuristic artificial general intelligence (AGI) 'post-singularity' should readily offer the capacity to action this; through agents who understand what consists of being human as humans understand it [2–4], these would require the additional precepts of being a human devoid of the knowledge of one's socio-economic status or political attitudes (medically equivalent to an individual with what the suggested term of socio-politico-economic self-amnesia). What is suggested here-in rather, is that the AI technology of the current era would be able to offer an algorithmic social contract that would be at a technological ideal without needing any 'self-amnesia' due to the fact that they would not be artificially generally intelligent and more functional as programmable tools. To generate such algorithms for societal fairness and justice, there is an inherent necessity to address the issues of fairness, inherent bias discrimination with current and future AI models. It has been demonstrated that based on psychological appraisals of individuals, a human non-AI approach would be subject to common cognitive–neural processes and risk taking that would detract from the fair decision-making required for the veil of ignorance and its Difference Principle [12].

There are lessons that can be derived from the enactment of a Rawlsian justice system by current and future AI to also translate to other schools of political philosophy by an artificially intelligent agent-based platform. The selection of artificially intelligent agents to make decisions has both practical and ethical considerations. Practical considerations include resource considerations, for example machine learning algorithms are already being utilized in clinical research settings to act as primary or second opinion readers of radiological images in a clinical setting such as cancer screening from breast mammography. The arguments for this fall on either increased diagnostic accuracy when compared to individual humans and those of non-inferiority to expert groups, but also one of speed, persistence (working 24 h per day) and those cost-efficacy supporting under-resourced and financially rationalized health systems [1, 9]. In a comparative way, artificial intelligent agents may be cheaper decision-makers due to their digital design, though as this General AI is not yet present, it is unlikely such systems will be cheap in the short-term. Applying an AI-based system to carry out a process of a Rawlsian "Veil of Ignorance' (likely via simulating individuals and a society 'in the future') and justifying choices based on different principles can offer a possible means to achieving social justice through fairness.

4 Artificial intelligence ethical themes in a social contract

The genesis of an algorithmic AI approach for a social contract has several challenges to overcome based on recent experience. For example, the ongoing criticisms of algorithmic decision-making systems (ADS) in widespread use such as in hiring, lending, judicial and legal decisions, housing, healthcare and education [17].

For an algorithmic social contract, the core needs of data objectivity in the current era would require a significant overhaul as existing socially responsible data practices are generally lacking; limited data to date is truly objective. As a result, a system serving an entire nation state would require mechanisms to overcome bias and support fairness across a multitude of disciplines beyond those of internal technical decisions presented here. Additionally, the concepts presented contain assumptions about data objectivity and algorithmic fairness, and also the ease with which one can model 'optimal life courses,' and the straightforwardness of their implementation, which in real-world practice would require extensive digital architectural overhauling hard and software innovation and population level behavioral interventions, all of which would be complex and resource heavy.

Considerations here for the application of big data for algorithmic AI approaches include the need to overcome the bias issues of AI discrimination, security, ethics and colonialism [17, 18]. Algorithms applied need to be: (i) safe for all affected by them, (ii) reliable and (iii) available for utility, (iv) achieve an appropriate balance between privacy and security, (v) be explainable (through black-box, white-box or constructivist approaches), (vi) allow transparency and visibility of the social consequences of AI algorithmic decision-making. (vii) be legal under the jurisdiction through which it exists, (viii) be ethical. They will need to be a (ix) socially considerate and responsible and achieve fairness by having bigger and better data: (x) devoid of individual or group engineered and designed biases, (xi) devoid of data sources from non-diverse origins, groups, socio-economic strata, gender and geographies, (xii) devoid of those created for the purposefully prejudicial application. (xiii) Favoring freedom over control (including informed and uninformed control), (xiv) allowing a free and fair trade-off for digital independence and dependence.

5 Solution through modelling life journeys

Life journeys can be mapped from life to death. These can be punctuated with development stages and appraised through checkpoints in life that can be considered necessary for a 'well-lived' or high-quality existence that can be considered a 'good life'. This life course theory approach can offer examples models of a good life such as the World Bank's five key transition approach [22], learning, going to work, staying healthy, forming families, and exercising citizenship. Here there is child dependency on adult independence, educational transition from primary to secondary and higher moving into the workforce and those transitioning into responsible and productive citizenship that in turn allows the achievement of becoming economically productive society members and having the best chances for well-being and good health.

Life course models of this type offer trajectories where quantifiable outcomes such as health risk, happiness and wellness can be measured. This in turn allows the use of quality-of-life assessment instruments to be applied that can capture specific outcomes such as disease burden. For example, A life trajectory can be drawn, and disease burden conveyed by the metric of disability-adjusted life years (DALYs) which measures overall disease burden through the sum of the number of years of life lost (YLL) due to ill-health, disability or early death and the years lived with disability (YLD). With a clear set of quantifiable life trajectory, it is possible to apply the mathematics of route theory and routing games in game theory to appraise life trajectories to select the most appropriate 'life route' behind the veil of ignorance.

For example, Pigou's route example [14, 19] can be modified there will one quantity of people in a society (quantity of traffic) who travel from A (birth) → B (death) via two routes. Route 1 ('wide highway') where no matter how much of society enter this route, they will have the same quality of life. Route 2 ('narrow highway') is considered a 'short cut' where there can be congestion due to resource sharing so the quality of life in this trajectory will depend on the proportion of society that enter this route. For example, if 75% of the population take Route 1, their DALY will all be 10 and for the 25% of the population that take Route B, their DALY will all be 5.

The flow (proportion) of society members in Route 1 can be f and the flow(proportion) of society members in Route 2 will be $1-f$. We can work out the cost of the flow:

$$C(f) = 1 \times f + (1-f)^2$$

To calculate the best way for this flow (Optimal flow or f^*) or society to distribute itself for these life trajectories, can be performed through the differentiation of the cost function to find where the differentiation is nil to highlight the optimal flow, which is at 0.5 (or half the traffic in each life route). So the best possible cost for distributing society by life trajectory would 50% in each life trajectory.

$$f *= \frac{1}{2}, C(f *) = 7.5 \text{ DALYS}$$

The average societal DALY would be 7.5. However at an individual level, without the veil of ignorance, everyone would take the route with the lowest DALY burden, taking Route 2, so that the Nash flow would be zero:

$$f = 0, C(f) = 1$$

Consequently, the general cost of a Nash flow with this routing game approach is at most 4/3 of the minimum-latency flow for two-node, two-link networks with non-negative linear latency functions, the "price of anarchy" (PoA) or the veil of ignorance benefit (VoIB) would be 33% more than an individual rational self-selecting ('selfish') approach and likely to increase with other routing constructs such as those with affine delay functions. Of interest, the veil of ignorance benefit (and the Price of Anarchy) remains consistent when more life trajectories are interconnected as this value of 4/3 benefits remains consistent in routing cases such Braess's paradox, when adding one or more roads (life trajectories) to a network can increase overall DALYs burden (slow down overall traffic flow) through it.

An extension of this approach follows that if a life path is mapped by DALYs or YLDs (Years of life with disability), then these can be minimized by identifying the shortest route or path through DALYs or YLDs. One approach would be to use a path optimization/adaptive routing process such as Dijkstra's algorithm (an extension of the A* approach) to identify the shortest path from a single source [23]. This system utilizes distance labels from a start node (s) to all other nodes on the graph with temporary nodes, and then iteratively calculates the shortest route by sequentially removing temporary nodes until the shortest path is clear. In concept, the reverse might be possible depending on the data so that the longest life journey can also be identified by an inverse Dijkstra-type approach (or an appropriate longest path approach matching human life journeys) depending on whether the distribution is NP complete.

Whilst the application of Dijkstra's algorithm here is purposefully a representative concept it is simplistic in light of necessarily complex real-world, real-life data with quantifiable outcomes of health risk, happiness and wellness. The data for such a proposal is not currently used for such an approach and may possibly be derived from existing sources such as insurance companies, governments, global and national surveys or future data sources amenable to such a pathway analysis.

6 Classical approaches

Once various life courses are selected 'behind a veil of ignorance', they can be pitted against each other to identify the most appropriate set of life courses acceptable in an ideal society and their social determinants to then develop a social contract. For this, a linear algebra approach to Rawls' Maximin has considered two opposing angles, for example, a Marxist one with progressive taxation and a 'trickle down' diametrically opposite one that enhances capital accumulation feedback, letting a small percentage of society possess most of the resources through a market economy (for example comparing two individuals, one who would feel base-fulfilled by having hard resources and another by having happiness by the progressive distribution of all goods). Based on classic liner algebra, where n unknown quantities are constrained by m relations and $m<n$, no unique solution can be achieved and there are no adequate constraints to achieve an agreement zone to set a range of rules for society [11]. To solve this, however, a machine learning approach can be introduced via an optimization machine learning algorithm such as Gradient Descent Ascent (GDA), where a machine learning model based on a convex function modified its parameters iteratively to minimize a given function to its local minimum [7]. Here various approaches of Batch, Stochastic and Mini-batch approaches can account for various intermediate and diametrically opposite states to select an optimal one behind the veil.

Additionally, decision theory with min-of-means can be hybridized to the ex-ante versus ex-post distinction of welfare economics to approach an answer to 'adversarial' choices before looking through the veil [16], or alternatively Groupwise Maximin Fair Allocation of Indivisible Goods [5].

With these mathematical approaches, there are several fundamental AI governance issues to address. These include (a) exactly who decides the most appropriate set of life courses acceptable in an ideal society? (b) Would it be selected by a group or council of human representatives or representative humans? Or (c) would humans have a veto for these decisions. These major issues of AI governance could also translate into the variability of algorithmic social contracts for distinct populations and their population choices. It also necessitates the need for any population under an algorithmic AI social contract to have the ability to choose and change the nature of any underlying algorithm and its applications, which should always offer the ability for population choice.

7 Algorithmic game theory approach

Rawl's originally presented the theory of Justice, highlighting that the concept he wanted to apply for a just distribution of resources had already been characterized in the economic and game-theory literature through Wald's Maximin process. Whilst this approach can then achieve societal resource growth, it ensures that if there are any segment of society to benefit, this cannot be at the expense of their forfeiting, which offers the ability to appraise this process as a zero-sum game where there should be no losers, only different levels of winning. Ultimately therefore for even the richest in society who may become richer, the worst-off must at the same time benefit too.

Such a concept can be applied to a dynamic game theory approach where a 2×2 grid could be formed by looking at the outcomes of the 'aggregate of society' compared to the 'worse- off'. In our model, if we assume in the game that there is a cut-off line from losses, then we can now also apply a Minimax approach to the Rawlsian system, by having a zero-sum game between the worst off and society to ensure a positive result for the worst off even when the rich thrive. This is because Von Neumann proved in 1928 that for any finite, two-player, zero-sum game the maximum value of the minimum (Maximin) expected gain for one player is equal to the minimum value of the maximum (Minimax) expected loss for the other [13, 24]. Here the Nash equilibrium each player receives a payoff is equal to both his maximin value and his minimax value.

In this situation, we can now apply a wealth of AI approaches to calculate a just society by comparing the outcomes and variables of both groups behind the veil. Here a mathematical saddle point could be created to choose lines to ensure no one entering society would fall below a certain standard or quality of life. The computer science approach to allow this tangible possibility would be through techniques such as alpha and beta pruning; here there is algorithm optimization to allow calculability by ignoring branches and trees in a decision game sequence that would have no extra decision value if there were to be explored.

Here, alpha is the highest value at any instance along the path of the Maximizer (initially – ∞) and beta is the lowest value at any instance along the path of the Minimizer (initially for alpha is + ∞), and so for pruning $\alpha > = \beta$. Each node records and updates alpha and beta values where alpha can be updated only when it's MAX's turn and beta when it's MIN's turn.

8 Overcoming biases and ensuring systemic fairness

In the case presented, if a nation state was to consider applying some of these machine learning (and more broadly AI) and algorithmic game theory approaches, then there would be certain steps necessary to initiate the programme. These include building a digital infrastructure for decision-making, collect high volumes of big data relevant to a national individual-level population with a wide range of individual variables, and prepare a methodology to apply the AI tools and appraise their outputs. It also requires a clear process to minimize the issues of data drift, false confounders, data and algorithmic bias and ensure algorithmic fairness.

Bias in artificial intelligence [15] derives from (1) Data Source: inherent bias in the data (domain bias) as a poor representation of reality (e.g. class imbalance, poorly labelled data, dataset shift), (2) Algorithm/Analytical: biases developed directly in the algorithm itself as not being adequately fit-for-purpose (correlation fallacies, overgeneralization, distribution shifts, hidden biases), and (3) User-based/ Subjectivity: those of confirmation (the prior assumptions of data collection for an intended application are flawed) (Fig. 1). Techniques applied to overcome these biases includes (i) better backbone convoluted neural networks, (ii) batch normalization, (iii) instance + batch normalization, (iv) data augmentation, mix match, (v) semi-supervised approaches including pseudo-labelling and domain adaption.

There are also a variety of metrics and frames to assess and ensure algorithmic fairness [6, 8] and these include those of (a) equalizing the odds of outcomes (using confusion matrices to minimize false positives and negatives), (b) equalizing the access to resources and opportunities between groups, (c) reducing unfairness by overcoming the unawareness of inherent lack of knowledge of unfairness between groups, and (d) overcoming demographic unfairness. Approaches to address the latter include (i) Accuracy equity, (ii) Conditional accuracy equity, (iii) equity of opportunity, (iv) disparate impact measurement, (v) counterfactuals, (vi) group vs individual fairness by overcoming traditional statistical group fairness measures such as outcome parity, error parity, decision boundaries by modifying individual probabilities, calibration, multi-calibration and scaffolding).

Together, these now also offer an approach to engage other theories from political philosophy through an algorithmic game theory and AI analytical approach. Here normative features and like-versus-like can identify what classifiers are beneficial to consider for each outcome. For example, other schools (non-Rawlsian) of egalitarianism

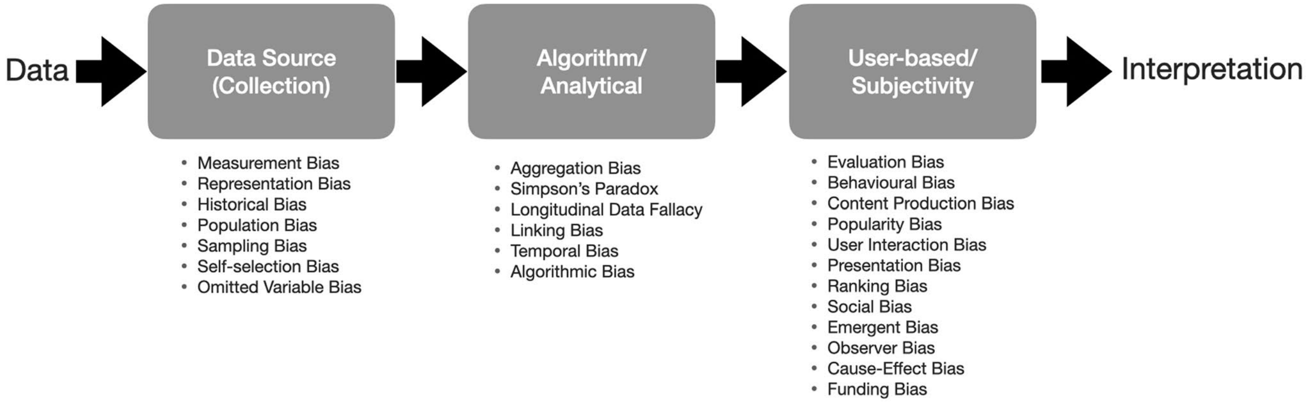

Fig. 1 Classification of Bias in Artificial Intelligence

can address structural disadvantage and injustice. Whilst it is considered that artificial intelligence ethics is incompatible with individual justice, other schools of egalitarianism may be bought forward that would be particularly amenable to AI, specifically those of egalitarian justice, which is a comparative notion between groups [6]. The valuation of redistribution comes from various schools such as welfare, resources, capabilities, political status, deontic justice, or even Waltzer's pluralist notion with spheres of egalitarian justice [6, 25]. Some cases here may need a trade off in selection (although these may not be appropriate for all cases such as a political election vote). Here, machine learning approaches may overcome luck egalitarianism by ensuring any selection for an individual's pathway is one based on personal decisions rather than inequalities that came from luck or those form deontic justice or those of distributive versus representational egalitarianism.

9 Conclusions

Developing a social contract has until now been considered to purely depend upon and be designed by its end users, humanity itself. Historically this has been subject to multiple irrationalities and biases inherent to human nature. Based on the objective capabilities of algorithms in decision-making with appropriate data and interpretation, the application of current and future algorithmic game theory and AI may offer a more measured selection of choices in a social contract for society and governmental policy; the potential here is fundamentally revolutionary. The large data lakes characterizing individuals within that society and the data linkages will be a key necessity for this innovation to take hold and will require data linkage with a purpose. These will transcend the classical push and pull of historical data linkage but rather have a more centralized approach with representative team-based information gathering, classification and presentation, building and presenting communities rather than technological functions. These algorithms will also still be prone to biases in the interpretation of their results will also need concomitant preemptive strategies to enhance algorithmic fairness and minimization bias.

For an algorithmic social contract, the core needs of data objectivity in the current era would require a significant overhaul as existing socially responsible data practices are generally lacking; limited data to date is truly objective. As a result, a system serving an entire nation state would require mechanisms to overcome bias and support fairness across a multitude of disciplines beyond those of internal technical decisions presented here. There will also need to be good and clear sources of human and population opinions and political choices. Additionally, the concepts presented contain assumptions about data objectivity and algorithmic fairness, and also the ease with which one can model 'optimal life courses,' and the straightforwardness of their implementation, which in real-world practice would require extensive digital architectural overhaling hard and software innovation and population level behavioral interventions, all of which would be complex and resource heavy.

If these factors can all be addressed and overcome with an appropriately safe digital infrastructure, applicable state data, protected with appropriate privacy, security, and ethical governance. AI capabilities of the current era render the application of these technologies for large-scale governance on populations technically uncertain and applying them would require considerable translation of computer science capabilities. There may be an opportunity to present population data to algorithmic game theory and AI tools and mandated to generate a social contract to do what humanity hasn't been able to achieve within itself, a fair, just and humane society that could be built on decisions from synthetic non-human elements, maybe in itself the highest level of humanity (*summa humanitas extrinsecus*).

Declarations

Conflict of interest HA is Chief Scientific Officer, Preemptive Medicine and Health Security at Flagship Pioneering.

References

1. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S.W., Karthikesalingam, A., King, D., et al.: Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ. Digit. Med. **4**(1), 65 (2021). https://doi.org/10.1038/s41746-021-00438-z
2. Ashrafian, H., Darzi, A., Athanasiou, T.: A novel modification of the turing test for artificial intelligence and robotics in healthcare. Int. J. Med. Robot. Comput. Assist. Surg. (2015). https://doi.org/10.1002/rcs.1570
3. Ashrafian, H.: Intelligent robots must uphold human rights. Nature **519**(7544), 391 (2015). https://doi.org/10.1038/519391a
4. Ashrafian, H.: Artificial intelligence and robot responsibilities: innovating beyond rights. Sci. Eng. Ethics **21**(2), 317–326 (2015). https://doi.org/10.1007/s11948-014-9541-0

5. Barman, S., Biswas, A., Krishnamurthy, S. K., Narahari, Y.: Groupwise Maximin fair allocation of indivisible goods. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). (2018) www.spliddit.org
6. Binns, R.: Fairness in machine learning: lessons from political philosophy. In: Proceedings of Machine Learning Research (vol. 81) (2018)
7. Daskalakis, C., Panageas, I.: The Limit points of (optimistic) gradient descent in min–max optimization. In: NeurIPS—32nd Annual Conference on Neural Information Processing Systems (2018)
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12 (pp. 214–226). New York: ACM Press. (2012) https://doi.org/10.1145/2090236.2090255
9. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
10. Harsanyi, J.C.: Can the maximin principle serve as a basis for morality? A critique of John Rawl's theory. Am. Polit. Sci. Rev. **69**(2), 594–606 (1975)
11. Houchmandzadeh, B.: Rawls's original position is not sucient to specify the rules of cooperations. HAL Open Science. (2018) https://hal.archives-ouvertes.fr/hal-01922792v2. Accessed 22 Mar 2022
12. Kameda, T., Inukai, K., Higuchi, S., Ogawa, A., Kim, H., Matsuda, T., Sakagami, M.: Rawlsian maximin rule operates as a common cognitive anchor in distributive justice and risky decisions. Proc. Natl. Acad. Sci. U.S.A. **113**(42), 11817–11822 (2016). https://doi.org/10.1073/pnas.1602641113
13. Kjeldsen, T.H.: John von Neumann's conception of the minimax theorem: a journey through different mathematical contexts. Arch. Hist. Exact Sci. **56**, 39–68 (2001)
14. Knight, V.A., Harper, P.R.: Selfish routing in public services. Eur. J. Oper. Res. **230**(1), 122–132 (2013). https://doi.org/10.1016/j.ejor.2013.04.003
15. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. arXiv. (2019) http://arxiv.org/abs/1908.09635
16. Mongin, P., Pivato, M.: Rawls's difference principle and maximin rule of allocation: a new analysis. Econ. Theor. **71**(4), 1499–1525 (2021). https://doi.org/10.1007/s00199-021-01344-x
17. Newell, S., Marabelli, M.: Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of "datification." SSRN Electron. J. (2015). https://doi.org/10.2139/ssrn.2644093
18. Parliament, E., for Parliamentary Research Services, D.-G., Castelluccia, C., le Métayer, D.: Understanding algorithmic decision-making: opportunities and challenges. Publications Office. (2019) https://doi.org/10.2861/536131
19. Pigou, A.C.: The economics of welfare. Transaction Publishers (1920)
20. Rawls, J.: A theory of justice. Harvard University Press (Belknap Press), Cambridge (1971)
21. Sniedovich, M.: Wald's maximin model: a treasure in disguise! J. Risk Finance **9**(3), 287–291 (2008). https://doi.org/10.1108/15265940810875603
22. The World Bank: World development report 2007: development and the next generation. The World Bank (2006)
23. Verscheure, L., Peyrodie, L., Makni, N., Betrouni, N., Maouche, S., & Vermandel, M.: Dijkstra's algorithm applied to 3D skeletonization of the brain vascular tree: evaluation and application to symbolic. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2010, 3081–4. (2010) https://doi.org/10.1109/IEMBS.2010.5626112
24. von Neumann, J.: Zur theorie der gesellschaftsspiele. Math. Ann. **100**, 295–320 (1928)
25. Walzer, M.: Spheres of justice: a defense of pluralism and equality. Basic Books, New York (1983)