V20

SHARING

ALIGNED VALUES, ARTIFICIAL INTELLIGENCE (AI), AND POLICYMAKING

ALIGNED VALUES, ARTIFICIAL INTELLIGENCE (AI), AND POLICYMAKING

Mashael Alzaid,
Data Scientist/Researcher, Saudi Arabia,
Mashaelalzaid@gmail.com

Dr. Manuel Schubert,
Managing Director at Behavia,Germany/Saudi Arabia,
manuel.schubert@behavia.de

Artificial Intelligence (AI) systems present significant opportunities and serious threats to the future of societal well-being. The G20 seeks to harness the benefits of AI for the good of all public services. This document discusses three major challenges with respect to the use of AI in public policy and presents a structured process that can contribute to a more fair and responsible values-centric use of AI technologies in public policy.

Global challenge

The Fourth Industrial Revolution is affecting societies around the globe. While it represents historic opportunities to improve quality of life and access to equal opportunities, it has also been the driver of inequality and the carrier of public harm (Larson et al 2016; Hannen 2020; Guo & Hao 2020). The increased use of Artificial Intelligence (AI) is both one the biggest opportunities and challenges for society. In response, the G20 have released the G20 AI Principles, stressing the need for responsible stewardship of trustworthy AI that reflects "human-centered values" (G20 2019). In continuation of these efforts, the G20 Digital Economy Task Force (DETF) is tasked this year with exploring opportunities to harness AI technologies for delivering more efficient and effective public services (G20 2021).

While we welcome the emphasis that the Italian Presidency places on using AI for public services, we see an urgent need to revamp policy principles and frameworks for the future use of AI in public policy. The more policymakers rely on AI as an enabler and accelerator of public services, the more likely we are to observe errors and flawed decision making. Three future challenges are identified with respect to the use of AI in policymaking: Biases, Responsibility, and Values.

1) Biases

One of the most notable examples of biased AI algorithms is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, used in USA courts to predict the probability that a defendant will become a recidivist. COMPAS has discriminated against black people, categorizing them too often as future offenders (Kozyreva et al 2021). Racial discrimination was also found in AI used in USA hospitals to predict which patients would need extra medical care (Shin 2020). These and other cases have sparked controversy, prompting authorities to disclose more information on data collection processes. Current practices, however, raise serious doubts about the validity of forecasts, possibly undermining the effectiveness and societal acceptance of AI-enhanced public services.

2) Responsibility

Another key challenge is related to the use of AI systems in specific policy contexts. First, AI algorithms are not yet able to "think" beyond data boundaries, and thus may overlook important linkages and interactions with other policy areas. Second, policymakers and citizens usually face difficulties in understanding the rationale of AI systems. For example, who will take responsibility for AI-enhanced services that are short-sighted, ill-designed, or discriminatory—the policymakers, the data scientists, or the AI itself? AI standards for public policy need to address these questions and prevent a diffusion of responsibility in these multi-stakeholder contexts.

ALIGNED VALUES, ARTIFICIAL INTELLIGENCE (AI), AND POLICYMAKING

3) Values
AI systems are known to offer effective decision support based on historical evidence. The question is whether historical data is always the best predictor for future behavior. Societal values are not necessarily constant, as witnessed during the COVID-19 pandemic and in the Fukushima nuclear disaster. Rapid shifts are problematic for AI algorithms and might even affect public attitudes towards AI-enhanced policies. Research finds that people tend to distrust AI algorithms after witnessing mistakes, even if the AI proves generally useful (Dietvorst et al 2015).

Global solution

Given the problems described above, we see the urgent need for collective efforts to address these challenges and lay the right foundations for an unbiased, responsible, and values-centric use of AI technologies in policymaking. However, most reports on the use of algorithmic systems in the public sector are still either descriptive or theoretical (Ada Lovelace Institute, AI Now Institute, and Open Government Partnership 2021). Thus far, only a few empirical studies have examined the impact and effectiveness of policy measures aimed at achieving "accountability" in specific contexts. Therefore, the following selection presents the most prominent approaches and frameworks, seeking to elicit the key elements of an AI policy design process that can help G20 policymakers operationalize the G20 AI Principles.

1. Citizen participation
Citizen participation is a widely used policy approach in traditional policy areas, such as: community development, urban planning, and public procurement. It incorporates a public dialogue in which citizens get involved at various stages of the policy cycle with the opportunity to influence assessments and decisions. In these areas, citizen participation is known to garner public support for planning decisions; enhance societal acceptance and trust towards policy measures; and nurture citizen engagement or community well-being (University of Oregon n.d.; Beck 2012; Behavia 2020). Likewise, in the context of AI technologies, citizens should take an active part in the policy-making process as representatives of the target group, for example, by monitoring and overseeing the evaluation and decision processes before an AI solution can be applied for public services.

2. Responsible Design Framework
A process which can complement the previously mentioned citizen participation approaches is the Responsible Design Framework (RDF,

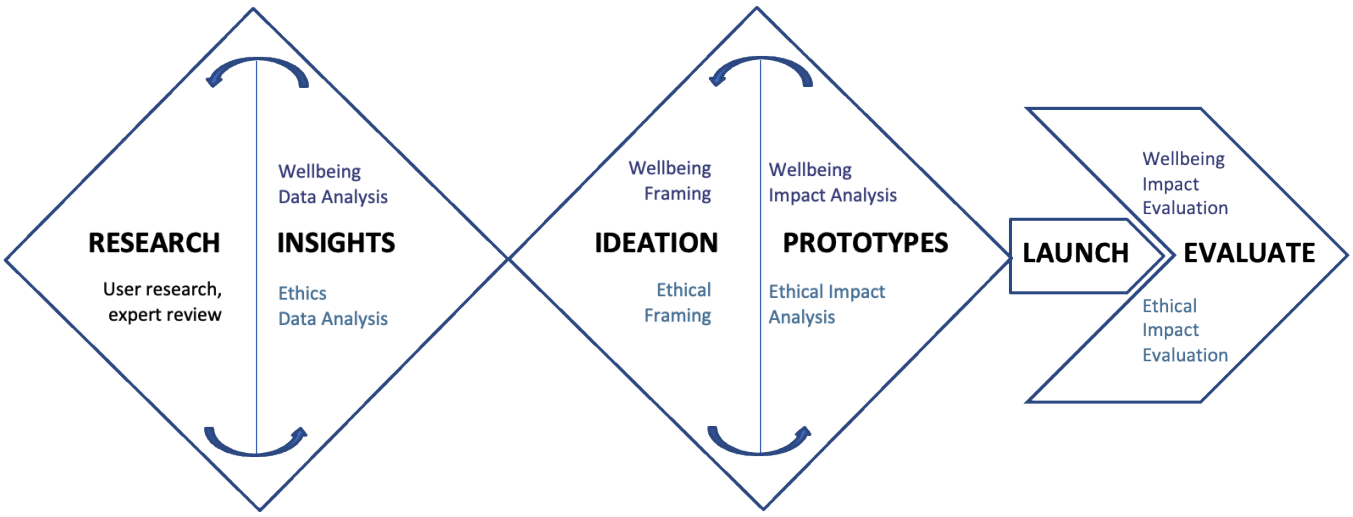

Fig 1: Responsible Design Framework (Source: Peters et al 2020: 37)

Peters et al 2020). The RDF is a process which focuses on ethical and well-being considerations in technology development by incorporating dedicated impact assessments at each stage of the engineering process.

3. Trustworthy Governance Structures

Enhanced governance structures for human-centered AI present another practical way to design reliable, safe, and trustworthy AI systems. Shneiderman (2020) identifies three core levels which address the key challenges discussed above from various angles:

1. Reliable Systems :This level suggests applying technical practices to software engineering teams that clarify human responsibility through audit trails and analysis tools. It also suggests adjusting software engineering workflows and supporting explainable user interfaces and verification and validation testing to enhance fairness and avoid harmful outcomes.

2. Safety Culture: The second layer encourages leadership commitment to safety through explicit statements about values, vision, and mission, and by making these visible to employees through frequent meetings. These meetings will be used to review and report failures and near misses, alignment with standards and best practices, and will provide safety training.

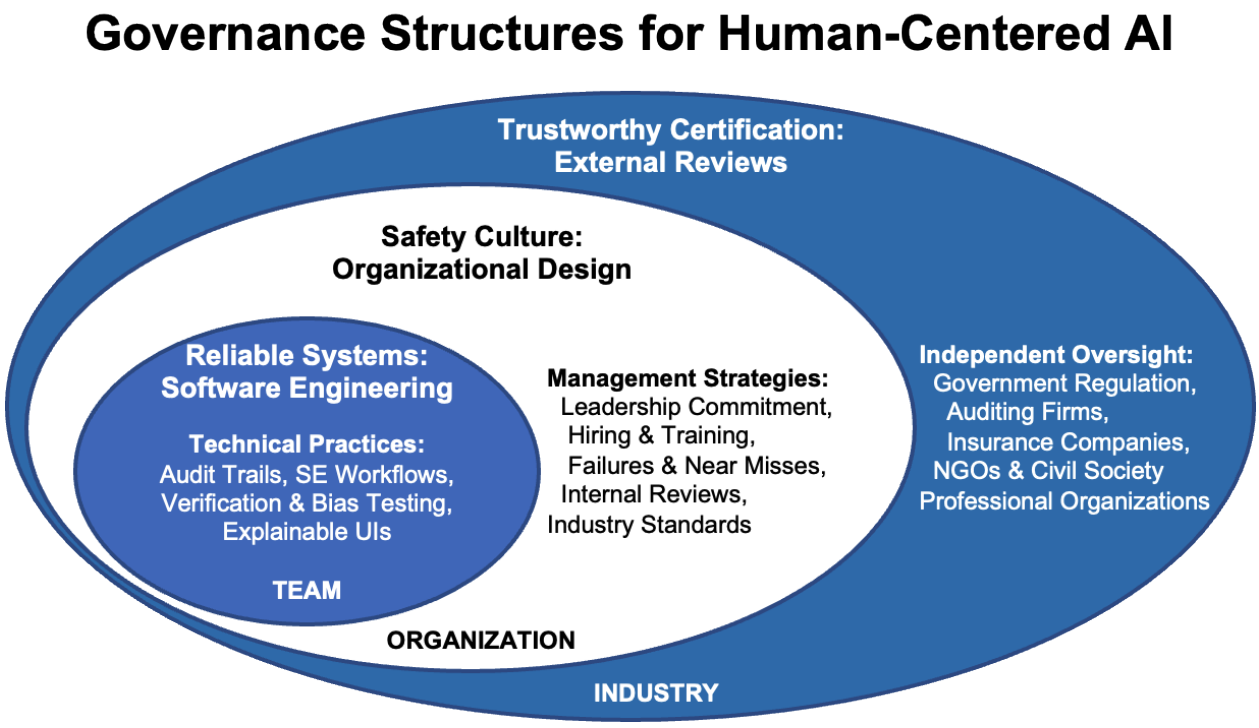

Fig 2: Governance structures for human-centered AI (Source: Shneiderman 2020: 3)

3. Trustworthy Certification: The final layer highlights the importance of independent oversight by external review organizations that increases the liability of the products and services.

Although the above structure (see Fig 2) was published recently, early evidence demonstrates the value of using flight data recorders, for instance, in making civil aviation safe—avoiding

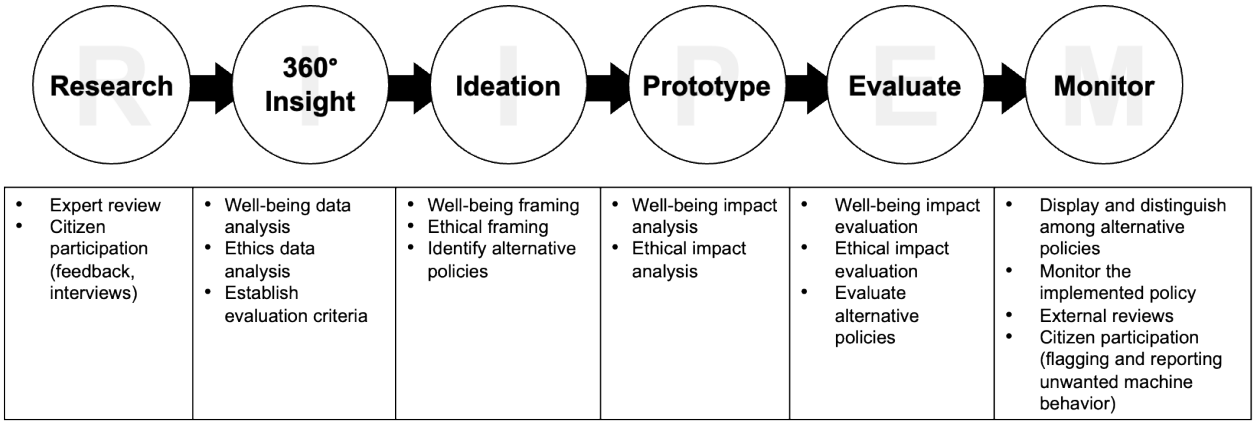

Fig 3: Policymakers' framework: People-oriented smart systems (Source: Authors' illustration)

accidents and improving training and equipment design (Grossi 1999).**4. An AI policy design process**

To operationalize the above frameworks and combine best practices from both a technical and a policy-making perspective, the following modified process is suggested to design policy in the context of trustworthy AI.

Research: The goal of this phase is to explore the specific needs of and possible barriers to the citizens who will be affected by the new technology, i.e., the actual target group. A key pillar of this phase is an in-depth assessment of specific service contexts. Typical approaches at this stage are not only reviews of relevant (grey) literature and qualitative assessments, but also collection of quantitative data through tracking systems, surveys, or user journey analysis.

360° Insights: This phase aims to identify the range and magnitude of risks, especially related to potential biases, ethical and accountability risks, as well as effects on perceived transparency and societal acceptance. This phase goes well beyond traditional risk assessments, as the harm inflicted on citizens needs to be forecasted based on (new) empirical evidence gathered during the previous stage.
Ideation: This phase comprises standard policy ideation and design-thinking formats to develop new policy solutions, explicitly considering the 360° insights derived during the previous phase.

The ideation stage should be supported by a simulation tool that flags risks at an early stage.

Ideation: This phase comprises standard policy ideation and design-thinking formats to develop new policy solutions, explicitly considering the 360° insights derived during the previous phase. The ideation stage should be supported by a simulation tool that flags risks at an early stage.

Prototypes: The ideated policy solutions are checked for feasibility by technical expert committees. Feasible solutions are operationalized as minimum viable products (MVP), e.g., policy prototypes. The prototypes are prioritized according to expected impact and costs, given budget and ethical considerations. At least two prototypes need to be selected to proceed to the next stage.

Evaluate: This phase involves experimentation or A/B testing on a random sample of the target group to evaluate the impact, costs, and possible harm caused by the prototypes. Preference should be given to real-world environments. Tracking systems should encompass the larger ecosystem to validate expected side effects identified during the 360° insights phase and to improve model fit.

Monitor: Upon completion of the evaluation phase, the most effective AI-enhanced service is scaled and continuously monitored to capture the long-term impact and side effects. The

ALIGNED VALUES, ARTIFICIAL INTELLIGENCE (AI), AND POLICYMAKING

ISSUE	LEAD QUESTIONS
RELIABILITY	What were the assumptions (i.e., will the system be used in a certain city, for a certain group/class of people) when the problem was defined? How was the data collected? What were the circumstances under which the data was collected? What are the backgrounds of the team members who collected the data? Does the data scientist team come from similar or different backgrounds? Is the data representative of the groups who will be using or impacted by the system?
OPENNESS	Was the system reproduced by an external team? And did they achieve the same results? Is society exposed to a portion of the data? Or was the public surveyed to support the results of the data? Can individuals in the target group get access to their data sets and analysis?
VALUE-ORIENTATION	What values are important to the society in the matter under study? Are these values reflected in the data according to blind review by a third party? Are the values reflected in the data in line with the G20 AI Principles?

Table 1: Smart Systems Assessment Checklist for policymakers (Source: Authors' illustration)

monitoring cadence might be altered over time due to biases that emerge from changing the use context (Friedman and Nissenbaum 1996), and close scrutiny is required if additional (AI-enhanced) services are launched that could interact with the existing one. This phase should also include external reviews and regular third-party auditing.

To complete the description of the AI policy design framework, we present a brief checklist below (see Table 1) that can help policymakers assess risks of AI systems and provide guidance during policy development stages outlined above.

The proposed approach can help policymakers ensure that the systems supporting their decisions are safe and human friendly. Nonetheless, policymakers may be limited in their access to data and unable to answer questions asked of them (i.e., owing to legal constraints). However, with greater awareness and more success stories of AI-enabled policymaking, these laws could be changed to support the involvement of policymakers.

ALIGNED VALUES, ARTIFICIAL INTELLIGENCE (AI), AND POLICYMAKING

Policy recommendations

Artificial Intelligence (AI) technologies offer a vast potential to foster well-being by improving equal access to opportunities and higher living standards. In recognition of these opportunities and in alignment with their objective to stimulate a transformative recovery through technological innovation, the G20 seeks to systematically harness digital technologies for more efficient and effective public services.

We welcome the strong emphasis the Italian Presidency places on supporting AI use in public services and more agile regulation by compiling the G20 Menu of Policy Options on productivity-enhancing digital transformation. However, due to the wide range of potential risks and damages that can be inflicted by AI-enhanced public services upon citizens, in particular minority groups, we also see an urgent need to provide detailed guidance and best practices to ensure a responsible, unbiased, and values-centric utilization of AI technologies in policymaking. We therefore call on the G20 to support international efforts to unify and integrate AI standards in the policy cycle.

In pursuing this goal, the G20 should take the following actions:

1. Task the G20 Digital Economy Task Force (DETF) with developing a policy toolkit with case studies on the responsible, unbiased, and human-centric use of AI technologies in policymaking.

2. Encourage technology partnerships between public and private sector organizations and research facilities to identify globally accepted AI standards.

3. In cooperation with the OECD, task the G20 Framework Working Group (FWG) with exploring new AI-compatible policy design processes and conducting regulatory impact assessments to get a clear understanding of the underlying assumptions and the effectiveness of new AI policy processes.

4. Expand the mandate of the G20 Behavioral Insights Knowledge Exchange Network (BIKEN) to consult and support the DETF and FWG in designing policy processes which incorporate mandatory checkpoints for ethical reviews in accordance with the ethics standards of behaviorally informed policy interventions.

References

Ada Lovelace Institute and Open Government Partnership. 2021. Algorithmic Accountability for the Public Sector. (August 23, 2021). https://www.opengovpartnership.org/documents/%20algorithmic-accountability-public-sector/

Beck, L. 2012. Anti-Corruption in Public Procurement: A Qualitative Research Design. Passau University. PhD diss. University of Passau.

Behavia. 2020. Behavioral Insights for City Planners and Architects. https://behavia.de/behavioral-urban-design/.

Dietvorst, B., J. P Simmons, and C. Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." Journal of Experimental Psychology: General 144(1): 114126-. http://dx.doi.org/10.1037/xge0000033

Friedman, B., and H. Nissenbaum. 1996. "Bias in Computer Systems." ACM Transactions on Information Systems 14(3): 330347-. https://doi.org/10.1145230538.230561/

G20. 2019. Ministerial Statement on Trade and Digital Economy. Annex. 1–14. https://www.mofa.go.jp/files/000486596.pdf

MITD (Ministro per l'innovacion tecnologica e la transizione digitale). 2021. First Digital Economy Task Force Meeting. https://innovazione.gov.it

ALIGNED VALUES, ARTIFICIAL INTELLIGENCE (AI), AND POLICYMAKING

G20. 2021. "The Digital Ministers Approves a Declaration Identifying 12 Actions to Accelerate the Digital Transition of the Economy and Governments."
https://www.g20.org/the-digital-ministers-approves-a-declaration-identifying-12-actions-to-accelerate-the-digital-transition-of-the-economy-and-governments.html

Grossi, D. 1999. "Aviation Recorder Overview." Proceedings of the International Symposium on Transportation Recorders, pp. 153–164.
http://iasa.com.au/folders/Publications/pdf_library/grossi.pdf

Guo, E., and K. Hao. 2020. "This Is the Stanford Vaccine Algorithm That Left Out Frontline Doctors." MIT Technology Review. (December 21, 2020).
https://www.technologyreview.com/20201015303/21/12//stanford-vaccine-algorithm/

Hannen, T. 2020. "What Went Wrong with the A-level Algorithm?" Financial Times.
https://www.ft.com/video/282ecd1f-84024-bf48-ee73-d179ce5fcc2.

Kozyreva, A., P. Lorenz-Spreen, R. Hertwig, S. Lewandowski, and S. Herzog. 2021. "Public Attitudes Towards Algorithmic Personalization and Use of Personal Data Online: Evidence from Germany, Great Britain, and the United States." Humanities and Social Science Communications 8(117): 111-.
https://www.nature.com/articles/s4159900787--021-w

Larson, J., and J. Angwin. May 23, 2016. "Machine Bias." ProPublica.
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Reidl, M., and B. Harrison. 2015. "Using Stories to Teach Human Values to Artificial Agents." Association for the Advancement of Artificial Intelligence: 18-.
https://www.cc.gatech.edu/~riedl/pubs/aaai-ethics16.pdf

Peters, D., K. Vold, D. Robinson, and R. Calvo. 2020. "Responsible AI: Two Frameworks for Ethical Design Practice." IEEEXplore 1(1): 3447-.
https://ieeexplore.ieee.org/document/9001063.

Shin, T. 2020. "Real-life Examples of Discriminating Artificial Intelligence." Towards Data Science.
https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

Shneiderman, Ben. 2020. "Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems." ACM Transactions on Interactive Intelligent System 10(4): 31 pages.
https://dl.acm.org/doi/pdf/10.11453419764/.

University of Oregon. n.d. "The Theory of Citizen Involvement."
https://pages.uoregon.edu/rgp/PPPM613/class10theory.htm

www.values20.org