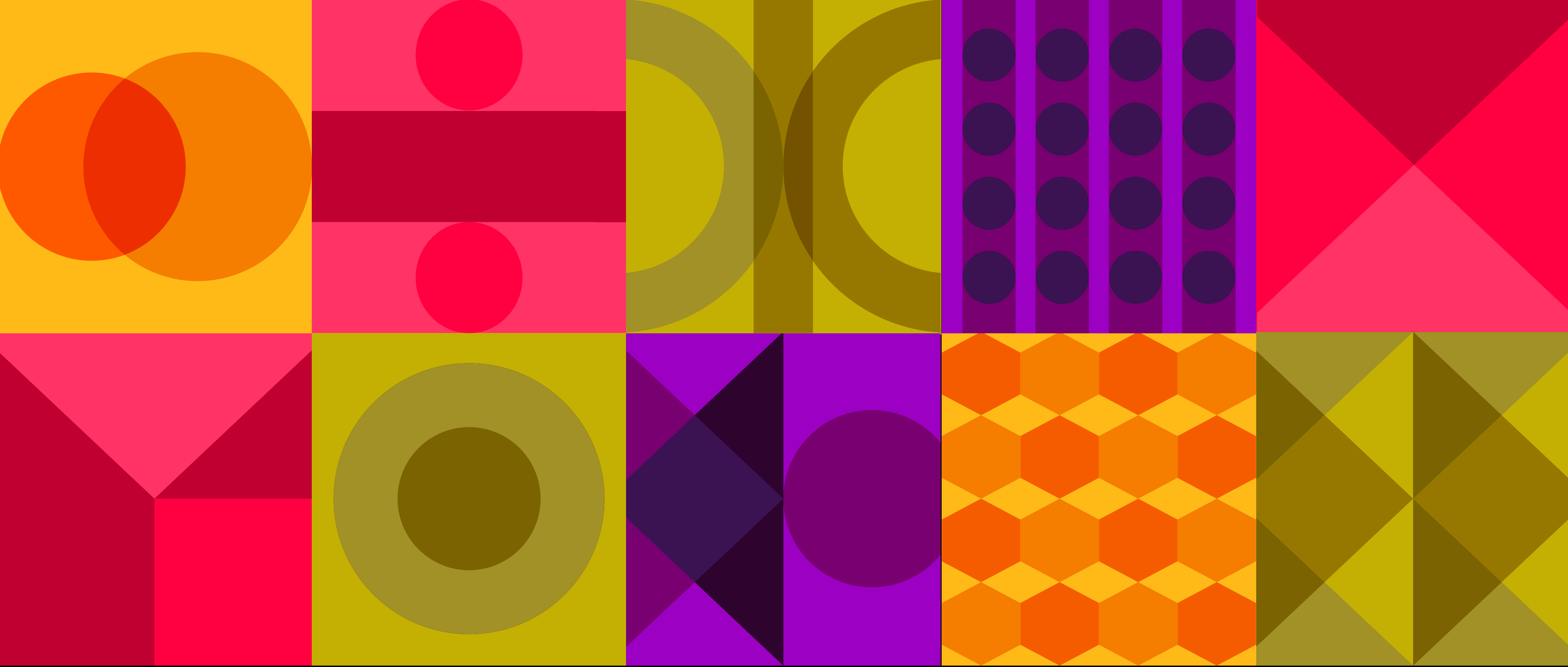


nesta

Combining Crowds and Machines

Experiments in collective intelligence design

AUTHORS

Eva Grobbink
Kathy Peach

ACKNOWLEDGEMENTS

This report is based on the findings from 12 collective intelligence experiments that were funded, but not conducted by Nesta. The experiments were carried out by:

Elías Fernández Domingos, Jelena Grujić and Tom Lenaerts (Artificial Intelligence Lab, Vrije Universiteit Brussel), Inês Terrucha (Universiteit Gent and Vrije Universiteit Brussel), Remi Suchon (Université Libre de Bruxelles), Francisco C. Santos (INESC-ID, Universidade de Lisboa), David Baltaxe, Greg Willcox and Louis Rosenberg (Unanimous AI), Mark Burgman and Alexandru Marcoci (Imperial College London), Colin Talbot (University of Cambridge and University of Manchester), Vito Trianni, Giulia Andrighetto and Eugenia Polizzi di Sorrentino (Istituto di Scienze e Tecnologie della Cognizione), Dario Albani (Sapienza University of Rome), Rachel Briscoe and Dan Barnard (Fast Familiar, formerly fanSHEN), Kris De Meyer (King's College London), Joe McAlister, John Coburn (Seven Stories), Daisy Tam (Hong Kong Baptist University), Tomas Holderness (formerly MIT Urban Risk Lab), Mart van de Ven (Droste), Natalie Widmann, Clément Habinshuti and Jaume Cardona (HURIDOCS), Jessica Dheere (Social Media Exchange – SMEX), Rob Procter, Maria Liakata (University of Warwick and Alan Turing Institute for Data Science and AI), Miguel Arana Catania (Alan Turing Institute for Data Science and AI), Nataliya Tkachenko and Yulan He (University of Warwick), Arkaitz Zubiaga (Queen Mary University of London), Karel Verhaeghe, Sébastien Hoorens, Antoine Pirard and Koen Gremmelprez (CitizenLab), Lal Chadeesingh and Emily Larson (Behavioural Insights Team), Simon Woodhead (Eedi), Yvonne McDermott Rees (Swansea University), Jeff Deutch (Syrian Archive/Yemeni Archive), Dearbhla Minogue (Global Legal Action Network GLAN), Enrique Piraces (Carnegie Mellon University), Adam Harvey (VFRAME), Neal Reeves and Elena Simperl (King's College London), Gopal Ramchurn, Briony Gray, Luca Capezzuto and Steven Ball (University of Southampton), Kobi Gal and Naama Dayan (University of Edinburgh), Darlene Cavalier (Arizona State University) and other team members that might not be listed here.

We would also like to acknowledge the many hours of work contributed by volunteers and citizen scientists without whom these experiments would not have been possible.

We are very grateful for the valuable feedback on this report from our colleagues at Nesta's Centre for Collective Intelligence Design, Peter Baeck and Aleks Berditchevskaia.

The illustrations for this report were created by Kirstin Smith. The layout for this report was designed by Mike Green of Green Doe Graphic Design.

ABOUT NESTA

Nesta is an innovation foundation. For us, innovation means turning bold ideas into reality and changing lives for the better.

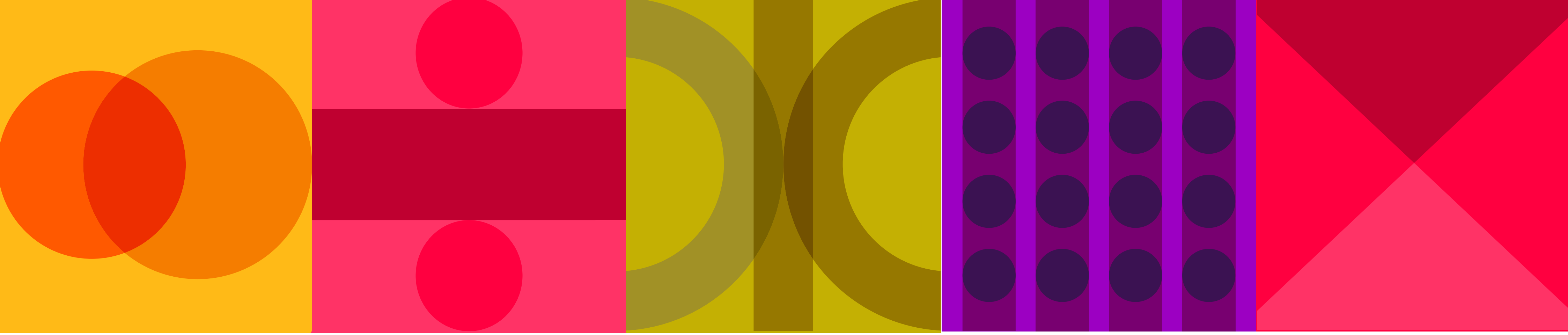
We use our expertise, skills and funding in areas where there are big challenges facing society.

Nesta is based in the UK and supported by a financial endowment. We work with partners around the globe to bring bold ideas to life to change the world for good.

To find out more visit **www.nesta.org.uk**

If you'd like this publication in an alternative format, such as Braille or large-print please contact us at: **information@nesta.org.uk**

Published June 2020



Combining Crowds and Machines

Experiments in collective intelligence design

nesta
Collective Intelligence

Executive summary

Tackling some of the most complex challenges of our time requires progress in how we think and act together. New technologies, including artificial intelligence (AI), allow us to mobilise human intelligence in new ways and at greater scale.[1]

At Nesta's Centre for Collective Intelligence Design, we have been focusing on advancing the knowledge and practical applications of collective intelligence in fields with public benefit, such as health, international development or digital democracy. Yet, in spite of the many emerging opportunities to use novel combinations of human and machine intelligence, we still know relatively little about what works and how to do it well.

Through our Collective Intelligence Grants Programme – the first fund of its kind – we supported 12 diverse organisations worldwide to conduct practical experiments that increase our understanding of how to make the most of the new technologies available to help with collective thinking and acting.

The experiments contribute new insights into how we can improve our decision-making, enable effective co-operation, make better use of citizen-generated data and increase the effectiveness of participation in collective intelligence initiatives.

Some of the findings provide the basis for further research, whereas others will be directly applicable in practice. For example, the experiment led by Swansea University will help pave the way for using crowdsourced evidence in a case pursuing accountability for war crimes in Yemen.

Seven key insights that we gained from the experiments

To make better collective decisions, delegate to AI

In collective risk dilemmas, groups of people must work together to reach a target, such as reducing CO_2 levels, or everyone suffers. But this requires individuals to make sacrifices, and levels of co-operation are often low. The Artificial Intelligence Lab at the Vrije Universiteit Brussel found that groups were more successful when people delegated responsibility to an AI autonomous agent to make decisions on their behalf. This was because people picked autonomous agents programmed to act in the interests of the collective, rather than those programmed to maximise benefit to the individual. Could it be that delegating to AI encouraged people to think longer term or reduced their fear of being cheated by other participants?

Want happier voters? Let them swarm!

Unanimous AI tested whether algorithms modelled on the swarm behaviour of honeybees could help politically polarised British voters to agree on government priorities. Unlike traditional voting, swarming allowed participants to see the groups' emerging consensus during the decision-making process and converge on a decision together in real time. The experiment found that voters were consistently happier with the results generated through swarming than those produced by majority vote – the most common decision-making method in modern democracies.

Use AI to stop people following the herd

In any collective decision-making process, group members are influenced by one another, which can reduce the accuracy of the decision. Researchers based at the Istituto di Scienze e Tecnologie della Cognizione (ISTC) found that mediating group decisions through an AI system reduced herding (where people go along with the majority view) and led to more accurate outcomes. The AI system pushed individual participants to continue exploring all the options for longer and stopped a group majority view being formed too early on in the decision-making process.

When fast action is needed, let the crowd self-organise

In time-critical scenarios, such as food rescue efforts, effective co-ordination among different actors is key. Researchers at Hong Kong Baptist University found that decentralising volunteer co-ordination through a collective intelligence platform led to a fourfold increase in food rescued. Allowing volunteers to see quantities of leftover bread at different bakeries across Hong Kong in real time enabled volunteers to optimise their own collection routes by adapting to a changing situation.

To make digital democracy work better, use AI to help people find similar ideas

Engagement rates on digital democracy platforms such as Consul are high, but very few citizen proposals reach the public support necessary to be considered by the local government. The Alan Turing Institute for Data Science and AI found that features enabled through natural language processing (NLP), such as automated tagging of proposals, significantly reduced the time it took for users to find similar proposals and other citizens advocating for similar ideas. Next we need to test whether this finding helps people to avoid duplication of efforts and increases the impact of participation on digital platforms.

Offering better rewards or more varied tasks doesn’t get better results from crowdworkers

In disaster situations, organisations increasingly rely on crowdsourcing data analysis from crowdworkers. Researchers at the University of Southampton found that higher pay for crowdworkers did not always result in more or higher-quality work and even had an adverse impact on labelling accuracy. In addition, crowdworkers favoured repetition over variation in tasks, as they could complete more tasks in less time, and were more likely to respond to feedback from other crowdworkers than experts.

AI recommendations can increase the engagement of citizen scientists by helping them discover less popular projects

Citizen science platforms host thousands of projects to which volunteers can contribute. Faced with a large number of options, it is difficult for individuals to find the project that best suits their preferences and skills. Researchers at the University of Edinburgh tested different recommendation algorithms on the citizen science platform SciStarter to better match users with projects they are interested in. They found that an algorithm that used a technique called matrix factorisation increased the activity of volunteers on the platform by recommending less well-known projects.

This report is based on the findings from 12 experiments that were funded, but not conducted, by Nesta. The experiments have shown the cutting edge of collective intelligence design and demonstrated the significant potential of crowd and machine collaboration to solve social challenges. Collective intelligence, however, is still a nascent area for research funding. We hope these experiments will encourage more funders to invest in collective intelligence and inspire more researchers and practitioners to help us advance it.

Introduction: Experiments in collective intelligence design

At its simplest, 'collective intelligence' can be understood as the enhanced capacity that is created when people work together, often with the help of technology, to mobilise a wider range of information, ideas and insights.[1]

It has been around for a long time, but new technologies are increasingly transforming what can be achieved through collective intelligence, either by connecting more of us together or by combining the complementary strengths of machine and human intelligence.[2]

The complex challenges we face today – from the climate crisis to political polarisation – are ripe candidates for collective intelligence. Yet despite much promising practice, and a growing academic field, we still know relatively little about what works in designing and applying collective intelligence.

Nesta's Collective Intelligence Grants Programme[i] supports practical experiments that generate insights and knowledge about how collective intelligence can be designed and applied. By funding experiments, our goal is to accelerate learning about what works.

During 2019/20, we funded 12 diverse experiments with up to £20,000 each.

What is in this report

This report is based on the insights and learnings from the first cohort of grantees. It summarises the experiments funded, highlights the main findings and outlines their relevance for collective intelligence practitioners and researchers.

The 12 experiments are divided into four categories. **Chapter 1** presents those experiments that looked into how we can improve collective decision-making. **Chapter 2** covers the experiments that explored how to facilitate more effective collaboration. **Chapter 3** provides insights on how to make better use of citizen-generated data. Finally, **Chapter 4** features experiments on how to increase the effectiveness of participation in collective intelligence initiatives. The report concludes with a call for more funding in the area of collective intelligence design.

Intended audience

This report is primarily aimed at practitioners and innovators who want to apply collective intelligence to address social challenges.

We hope, however, that the insights will also inspire more funders, decision-makers and academics to take this research further.

[i] To find out more about the Collective Intelligence Grants Programme, visit https://www.nesta.org.uk/ project/collective-intelligence-grants

01

How to improve the quality of group decision-making

Many of the complex challenges we face, from ageing populations to air pollution, need collective agreement and action to solve them. They may need individuals to change behaviour and give up resources or convenience to benefit others. These trade-offs are hard enough, but as societies become increasingly polarised, finding common ground becomes both more difficult and critical.

Delegate responsibility for decisions to AI

Experiment 1: Will autonomous agents help people co-operate better?

Grantee: Artificial Intelligence Lab, Vrije Universiteit Brussel, in collaboration with INESC-ID, Universidade de Lisboa.

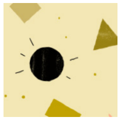
Key finding: Delegating decisions to autonomous AI agents improved group success, but people still said they'd prefer to remain in control.

Who is this relevant for?

- Digital democracy platforms.
- Designers of decision-support systems.
- Policymakers.

Overview

The Artificial Intelligence Lab at the Vrije Universiteit Brussel tested whether using AI improves co-operation and decision-making in collective risk situations. In such situations, the group must reach a consensus on a common challenge, such as the climate crisis, to avoid detrimental outcomes for everyone. The experiment results show that delegating responsibility for actions to autonomous agents increased co-operation and co-ordination within the group.



From the climate crisis to global pandemics, some of the most complex challenges we face are characterised by the presence of a collective risk. Often referred to as collective risk dilemmas, these challenges require individuals to make sacrifices and contribute to reach a collective target – such as reducing CO_2 levels. If the group fails, everybody suffers.[3] Behavioural experiments have shown that people only start to collaborate to avoid these disasters when the perceived risk is very high. Even then, co-operation levels are low, and the collective target is not often achieved.[3, 4]

This experiment tested whether autonomous agents could help increase co-operation between people in a collective risk dilemma. The experiment set out to understand how people would respond to the presence of autonomous agents and whether it would increase the collective success of the group.

What is an autonomous agent?

Autonomous agents are software programmes that can act autonomously on our behalf. They are able to respond to states and events in their environment independent from direct instruction by a human being. Agent software can range from simple programmes composed of a small number of rules to large and complex systems.[5, 6] When individual agents react to their local environment and other agents, they produce emergent collective behaviour.[2]

The researchers found that **co-operation increased when people delegated responsibility for making a decision to autonomous agents** (as shown by the results for delegation and customisation in Figure 1).

In the experiment, each participant chose one of several autonomous agents that either acted in the interest of the group (contributing to avoiding the disaster) or in the interest of the individual (trying to maximise the benefit for the participant). Previous research has shown that people, when delegating to other humans, tend to choose those who act more selfishly.[7, 8] In contrast, observations throughout this experiment suggested that people tended to delegate their decisions to agents that benefited the collective. This tendency has also been observed in previous experiments with autonomous agents.[5]

One explanation could be that people are forced to think longer term when delegating their decision-making or that acting through autonomous agents reduces the fear of being betrayed by others.[9] Despite not knowing which agent the other group members chose, the experiment participants could perhaps also rely on the fact that agents don't have the ability to cheat – they are programmed and thus unable to change their behaviour.

Despite these findings, a follow-up survey found that the majority of participants would still prefer to play the game themselves rather than delegating to an artificial agent. However, people reported they would be more willing to delegate their actions if they were able to customise the behaviour of the autonomous agents themselves, and not just choose from the five available agents (Figure 2).

Designers of future hybrid human and AI decision-making systems should therefore explore what the optimal level of control and agency is for people over autonomous agents.

Figure 1: Percentage of groups in each experimental treatment that reached the collective target through co-operation and 'avoided disaster'

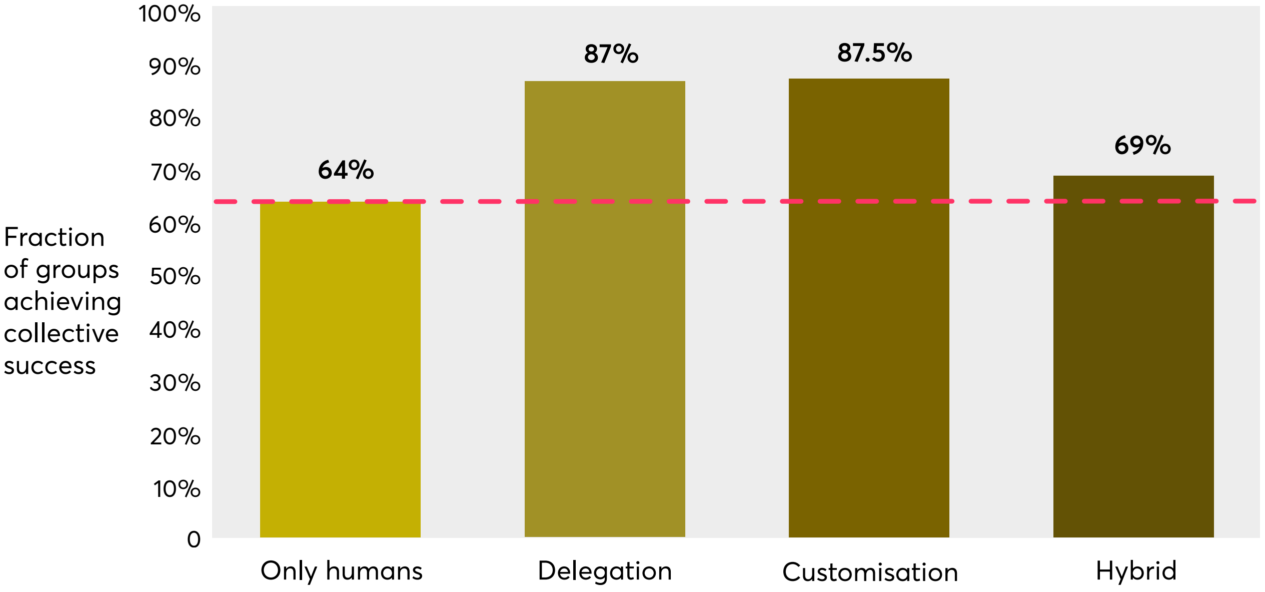
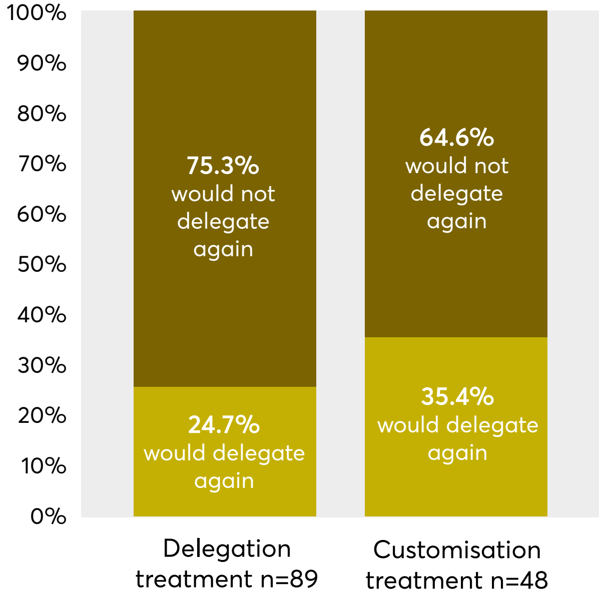


Figure 2: Percentage of participants that would choose to delegate or customise an agent if given the option again



How was this tested?

The lab-based collective risk experiment was run as a game of 10 rounds in groups of six. In each round, each player had to invest either zero, two or four experimental monetary units (EMUs) from a limited personal endowment (40 EMUs) into a collective pot. If at the end of the game the joint contribution of all members was above a threshold of 120 EMUs, they could keep the remainder of their personal endowment. If it failed to reach 120 EMUs, there would be a 90 per cent risk of a catastrophe occurring, which in turn would reduce the reward of all group members to zero. This means that, while participants made individual decisions to contribute, their choices always resulted in a collective outcome.

The researchers tested three different versions of the game where people either **delegated** their actions to artificial agents, **customised** their own agent, or played together with artificial agents in a **hybrid** group:

- **Treatment 1 – Delegate**. Participants chose one of five types of behavioural agents to play the game in their place. Three of the agents displayed unconditional behaviours. These agents played zero, two or four EMUs respectively in each round, until the group achieved the required threshold or the game ended. The remaining two agents displayed conditional behaviours: one of them only invested if the other group members invested below or equal to two EMUs in the previous round. The final agent only contributed if the rest of the group donated two or more EMUs in the previous round.
- **Treatment 2 – Customise**. Participants were asked to configure a conditional artificial agent that would act in their place during the experiment. This meant that participants had to determine the agent's initial action (in the first round), and the actions the agent would take when the group contributions were above, equal to or below a certain threshold in each round.
- **Treatment 3 – Hybrid**. Three humans and three artificial agents formed the groups. Only agent behaviours that were part of successful groups in Treatment 2 were selected – this represented the 'best-case' scenario, where humans were mixed with the best-performing agents. The participants did not know which of the other group members were agents and which were humans.
- **Control group – Humans only**. As the control treatment, the same experiment was conducted with six human players only.

In total, the experiments involved 186 people.

Use AI to mediate collective decision-making processes

Experiment 2: Will Swarm AI help politically polarised groups come to more collectively acceptable decisions than traditional voting methods?

Grantee: Unanimous AI, in collaboration with Imperial College London and Prof Colin Talbot.

Key finding: Politically polarised British voters were happier with decisions made through 'swarming' and Borda count methods; majority voting consistently produced the least satisfactory decisions.

Who is this relevant for?

- Governments and local authorities.
- Digital democracy platforms.
- Anyone looking to design better ways of enabling groups of people to make decisions on polarising topics.

Overview

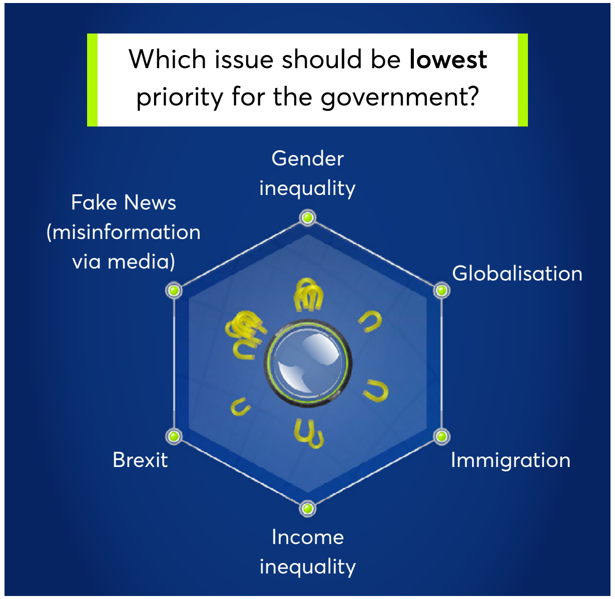
Unanimous AI tested whether algorithms modelled on the swarm behaviour of honeybees would enable politically polarised British voters to set government priorities more satisfactorily than traditional voting methods. They found that voters were consistently happier with the results generated through swarming than those produced by majority vote.

Polarisation is shaking societies across the world, from new democracies to long-established ones. In the US, voters have been increasingly sorted into two partisan identities. For much of 2018 and 2019, the UK public was polarised along Brexit dividing lines, as support for political parties fragmented and traditional allegiances were cast aside.

This experiment tested whether a decision-making platform that uses swarm algorithms could enable politically polarised British voters to rank different lists of government priorities[ii] in a way that led to more satisfactory outcomes than traditional voting methods.

[ii] For instance, one of the lists participants were asked to order was 'Rank the following UK Government objectives in order of their importance: Address climate change; drive economic growth; fix immigration policy; reduce crime; reduce poverty, solve the housing crisis'.

Participants answer questions as a 'swarming system' by collaboratively moving a graphical puck to select from a set of alternatives (each participant uses a mouse or touchscreen to manipulate one graphical magnet)



Which issue should be **lowest** priority for the government?

Unlike traditional voting methods, swarming allows participants to see the group's emerging consensus during the decision-making process and converge on a decision together in real time.[10] A total of 237 participants were recruited – a mix of Labour and Conservative supporters with split views on Brexit. In the experiment, participants were asked to rank lists of government priorities individually through a survey and together on the Swarm AI platform.

The ranked lists of priorities were then put to a control group of Labour and Conservative supporters who had not been involved in the process.

The researchers compared participants' satisfaction of results produced through swarming with satisfaction of results produced through the traditional voting methods of majority vote and Borda count.[iii, iv]

What is artificial swarm intelligence and how does it work?

Artificial swarm intelligence is a method that enables networked human groups to deliberate in real time. AI algorithms moderate the individuals' interaction as they decide between a set of options.

Unanimous AI's platform Swarm AI is a rare example of distributed AI and human groups working together on a task in real time. All participants log into the online platform at the same time. Individuals connect with each other and AI agents to form a closed-loop system where both the machine and individuals can react based on the behaviour displayed by others to change or maintain their preference. The algorithms are trained on data about behavioural dynamics of groups, not on the subjects they are debating. In a second step, a neural network model trained with supervised machine learning uses the interaction dynamics of the participants to generate a conviction index. This index estimates the group's confidence in the final outcome.[2, 10]

[iii] To analyse results through the majority algorithm, the objectives were ordered by the number of participants that ranked each objective as the 'most important' and ties were broken randomly.

[iv] Borda count is a simple method of combining rankings. In the experiment, using this method meant that each participant's ranking was converted into a score for each objective: one point for the 'most important' objective, two points for the second most important objective, three points for the third most important, etc., and the sum of these points across all participants in the group was calculated for each objective. The objectives were ordered from fewest points (most important) to most points (least important), with ties broken randomly.[10]

This experiment found that people involved in the decision-making process (the 'in-group') and a separate control group of people who weren't (the 'out-group') were **happier with the results generated through swarming than those produced by majority vote**. However, swarming and Borda count were perceived as producing similarly satisfactory results overall (as shown in Figure 3a and 3b).

The **majority voting method regularly produced the least satisfying outcomes** for both the in-group and the out-group. As majority vote is the most common method of aggregating voters' preferences in modern democracies, the results should give some pause for thought, and encourage further research into alternative and AI-mediated voting methods.

Figure 3a: Average satisfaction ranking (on a scale from 1 to 3, with 1 being the outcome people are most satisfied with) of in-group, by question number

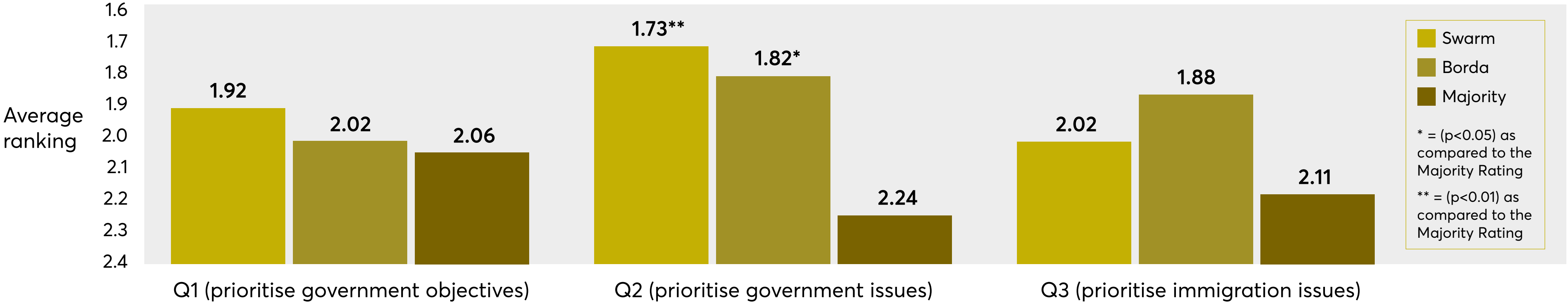
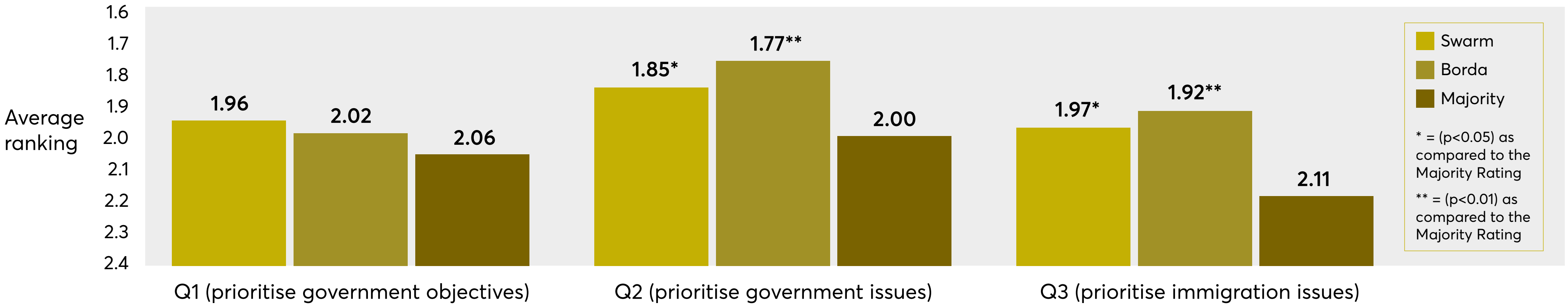


Figure 3b: Average satisfaction ranking (on a scale from 1 to 3, with 1 being the outcome people are most satisfied with) of out-group, by question number



How was this tested?

The researchers conducted four rounds of the experiment with groups of between 8-20 participants, consisting of male and female English citizens who were Labour or Conservative supporters with different opinions on Brexit. Participants were first asked to rank by preference the items on each list of government priorities through a survey, before ordering them through a swarming session.

The lists of government priorities were identified as politically divisive through pilot studies, where divisiveness was measured as the average difference between Labour and Conservative supporters' rankings of the different items.

In the swarming session, the groups started by selecting the least important objective out of the six items listed, then this item was eliminated from consideration and the group repeated the process, until there were two items left. For the final elimination, the group was asked which of the remaining items was the more important.

These four groups were considered the in-group, as they contributed to the ranked lists that they later scored for satisfaction. One out-group of 170 participants was also convened: they did not contribute to prioritising the list items, but only ranked their satisfaction with the lists generated by the in-group.

While the in-group participants were completing the ordering tasks on the platform, their survey results were analysed using majority vote and Borda count to generate two ordered lists. After the swarm session, in-group participants were redirected to a follow-up survey where they ranked the three ordered lists generated through majority vote, the Borda count method and the swarm session, based on how happy they were with each list. The satisfaction of the out-group with each of the in-group's rankings was also measured using a survey.

Experiment 3: Can multi-agent AI systems help us make better decisions by balancing out our biases?

Grantee: Istituto di Scienze e Tecnologie della Cognizione (ISTC) at the Italian National Research Council.

Key finding: Mediating group decisions through an AI system reduced herding – the tendency of people to go along with the group majority – and led to more accurate outcomes.

Who is this relevant for?

- Practitioners and researchers designing collective intelligence platforms and decision-making technologies.

Overview

The ISTC explored whether AI agents in a multi-agent system (MAS) could mitigate the negative impact of social influence in group decision-making. Their findings suggest that a MAS could reduce the effect of herding.

In any collective decision-making process, group members are influenced by each other. Herding – the tendency of people to go along with the group majority – is an example of social bias that can negatively impact collective decision-making.[11, 12] It means that groups can come to the wrong conclusion too quickly. In fields like medical diagnostics, the consequences can be serious.

This lab-based experiment explored whether the complexity of a problem affects the impact of social influence, and how much social information is actually useful for the accuracy of results. It then tested whether an AI system in the form of a MAS would be able to mitigate the negative impact of social bias.

What is a MAS and what did it do in this experiment?

A MAS consists of individual agents (typically software programmes) interacting with each other and the environment based on predetermined rules and constraints. The constant interactions and continuous learning by collecting and processing information allow a MAS to tackle problems that are difficult or impossible for an individual autonomous agent to solve.[13]

In ISTC's experiment, agents collected information about the participants' choices for the 'best' option. Based on this information, each agent decided whether it wanted to endorse one option or none (agents could also be 'undecided'). The number of agents endorsing an option represented how likely the MAS considered an option to be correct. The behaviour of the MAS changed over the course of the experiment. Initially, the MAS was programmed to give equal weighting to all choices, which encouraged participants to continue exploring and considering all the options. As the experiment progressed, the MAS was programmed to help the group achieve consensus. To do this, the MAS gave greater weighting to options that had been chosen more often by participants.

The researchers developed different conditions to simulate the varying availability of social information, in this case, knowing about each other's choices for 'best option', and the average group ratings of different options. In one of the conditions, participants interacted with a MAS, which provided information on options favoured by participants. Initially the MAS encouraged participants to explore all options equally by weighing them the same. As the experiment progressed, it encouraged consensus by giving greater weighting to options with more support from participants.

In the control condition, the group had to make a decision without sharing any information at all. Participants were randomly assigned to those different conditions to solve two tasks of different complexity.

The experiment found that:

- When confronted with the simple task, groups who had seen the choices of other participants identified the best option more often than individuals trying to solve the problem for themselves (as shown in Figure 4a).
- When given the more complex task, groups who were aware of others' opinions performed worse than people tasked with identifying the right option individually (as shown in Figure 4b).

This could imply that, **when faced with more difficult problems, people tend to trust the judgement of others excessively**, leading to a detrimental herding effect.

The results also found that more information increased group accuracy on the complex task, but decreased it on the simple one. When solving the simple task, the groups who knew only each other's best options (choice treatment) performed better than those who also were informed about the average group ratings (rating treatment) (Figure 4a). When the task was more complex, the groups with information about each other's best options and the group ratings identified the correct answer more often than those who could only see others' best options (Figure 4b).

The experiment found that the group decisions mediated through a MAS were more accurate on both tasks. In this experiment, the MAS helped to **reduce the negative effect of herding** by slowing down the process of decision-making and encouraging individual participants to continue exploring all the options. The preferences of the group were shared back and amplified by the MAS to help promote consensus only after individuals had explored the available options more widely. In this way, the MAS prevented participants from being overly influenced by others and stopped a group majority view being formed too early on in the decision-making process.

Using MASs to improve the accuracy of collective decisions may not be immediately applicable in the real world, but this is a promising area deserving of more research.

Figure 4a: Success rates for the 'simple' numerical sequences game

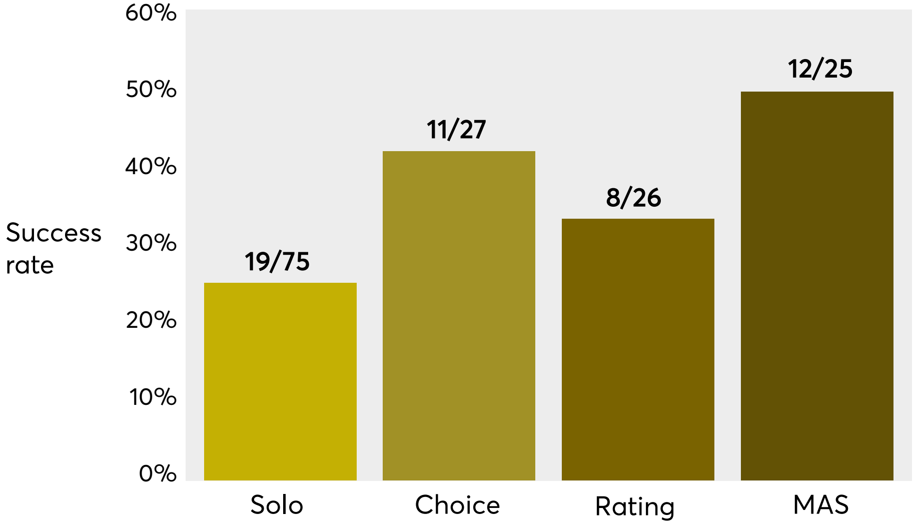
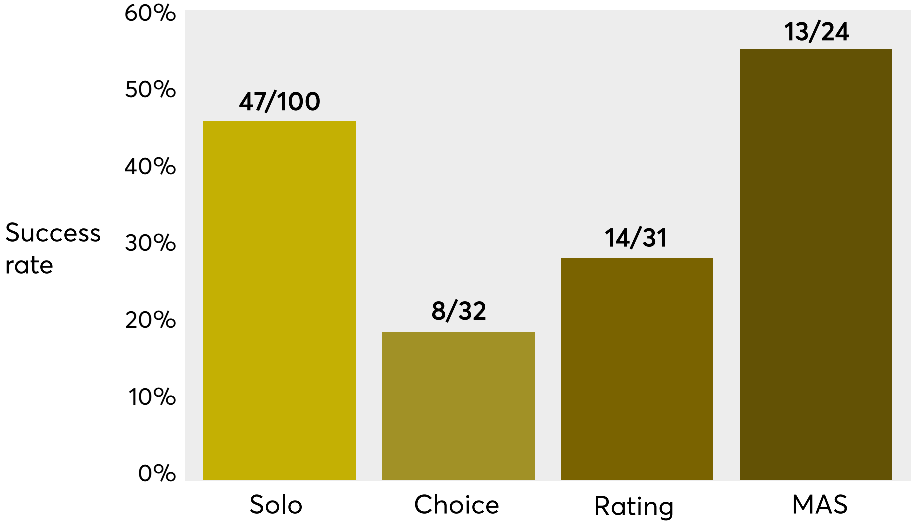


Figure 4b: Success rates for the 'complex' tuna and sharks game



How was this tested?

In the experiment, participants played two games of different complexity, where the correct option had to be chosen from several provided.

In the first game, participants had to decide which of two numerical sequences returned the highest values. This is not a particularly difficult problem. However, the sequencing of the intervals shown to participants was random. People might have only been shown part of an interval, which could differ each time the participant explored an option. This meant that some participants may have received only partial, and sometimes misleading, information.

In the second game, participants were shown images containing sharks and tunas and had to decide which one contained the most sharks. Each image was only shown for a couple of seconds. This task was much more complex than the first one, but there was no misleading information involved.

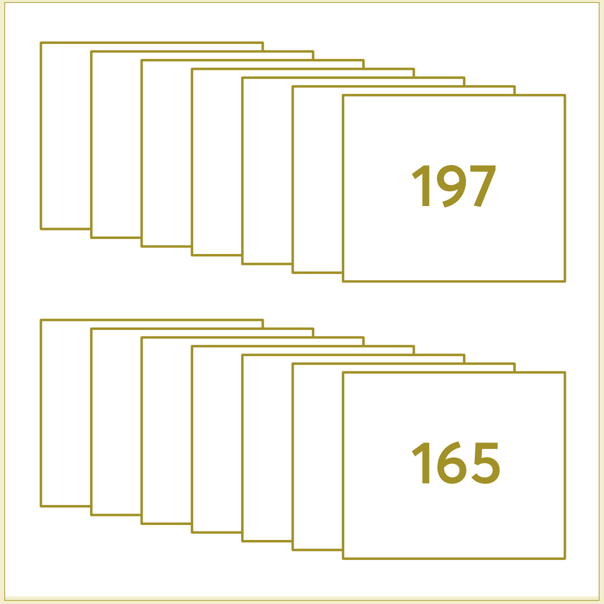
Each game consisted of several rounds. In each round, participants could explore one of the given options. Each participant was given 80 points. After the first 10 rounds, each round cost them 2 points if participants decided to acquire more information by exploring any of the options. During each round, participants rated each option on a scale from 1-5 and indicated which of the options they thought was the 'best' one. Participants could revise their previous judgement at any time during the game.

If at least 75 per cent of the participants agreed on the best option and if that group decision was correct, all participants retained their residual points; otherwise everybody lost everything.

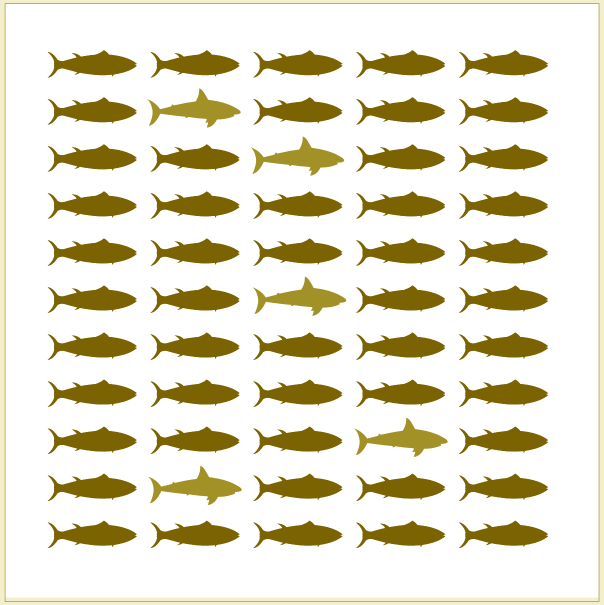
Participants were randomly assigned to one of the following treatments:

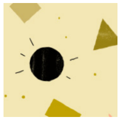
- **Solo** treatment, where people made decisions individually.
- **Choice** treatment, where five participants played together. They saw the **percentage of people that had chosen an option as their 'current best'**. This meant that they could also see whether they reached the quorum of 75 per cent for a certain option.
- **Rating** treatment, where in addition to the information provided in the choice treatment, participants were also able to see the **average group rating of an option on a scale from 1-5**.
- **MAS** treatment, **where participants interacted with the MAS**. The set-up resembled the choice treatment.

'Simple' numerical sequencing



'Complex' tuna and sharks





Improve metacognition and perspective-taking skills

Experiment 4: Can immersive digital storytelling help improve the ability of young people to see other perspectives?

Grantee: Fast Familiar (formerly fanSHEN) and Seven Stories, the National Centre for Children's Books.

Key finding: In all groups, participating students misidentified the feelings of characters in the story, suggesting an underdevelopment of emotional perspective-taking. After collectively debriefing on the experience, 84 per cent of participants reported that the experiment had helped them to take other perspectives.

Who is this relevant for?

- Teachers, counsellors, mentors, coaches or mental health professionals working with young people.
- HR professionals in companies that want to provide new forms of training to employees.
- Professionals in any setting where groups of people have to come to a decision through deliberation.

Overview

Theatre company Fast Familiar, together with neuroscientist Kris De Meyer, developed an experiential learning intervention to foster early adolescents' perspective-taking abilities. Students reported that the experiment helped them to think more about taking other peoples' perspectives.

Experiments 2 and 3 above demonstrated how AI could support us to find consensus in groups. But making the right decisions together also requires skills that are uniquely human: understanding different perspectives, being able to reflect on our own opinions and disagreeing productively. Research has found that reading stories in depth plays an important role in strengthening people's ability to understand different perspectives.[14] Increased digital media consumption has led to a shift towards superficial skim-reading, rather than becoming immersed in a story in depth. This is thought to be counterproductive for the development of empathy and perspective-taking skills.[15]

Together with a group of young adults and the organisation Seven Stories, Fast Familiar co-created *If I Were You*. The intervention combines digital storytelling and small-group discussions. The experiment was designed to test whether blending these different methods could foster students' capacity for perspective-taking.

How was this tested?

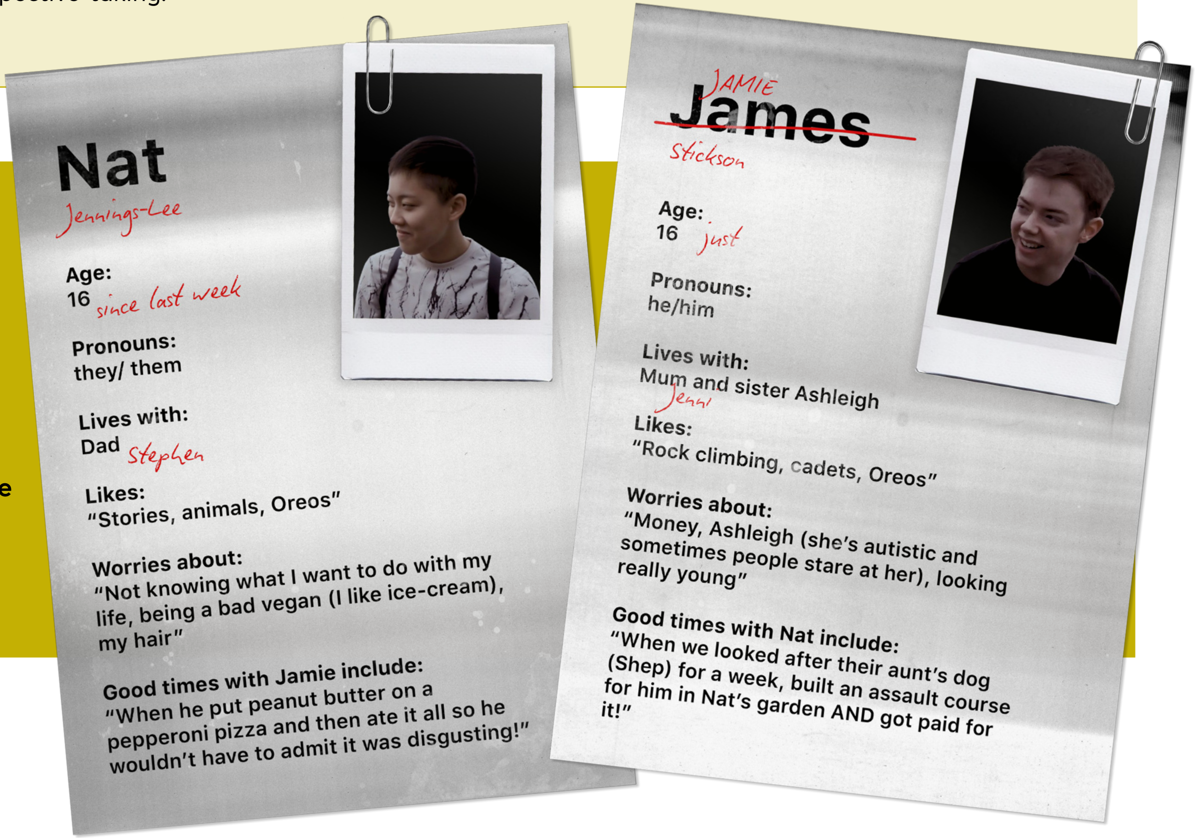
Psychology and neuroscience research has shown that perspective-taking skills can be fostered through reading fictional novels, where readers immerse themselves in the inner lives of different characters,[16] and through collaborative social reasoning (group deliberation).[17] And while digital technologies are seen as one of the drivers of the decline of deep reading, they also offer a potential solution. *If I Were You* is a playable digital novel for young adults, set against the backdrop of the climate emergency. Seven sessions were run in schools with students aged 13–17 in the north-east of England, and one session was run with adults.

During the sessions, two groups of people played in different rooms. Following the same story on tablets but from different characters' points of view, they received asymmetric information about situations and third characters via a bespoke control system implemented by the researchers. Over the course of the session, both groups were prompted with questions about what their own character was thinking or feeling, and how their character thought another person in the story was thinking or feeling.[v] Those questions were built into the experience to simultaneously trigger and measure cognitive and emotional perspective-taking.

The groups were also regularly asked to take decisions on what their character should do next. Group decisions were shared between the rooms, affecting the development of the story and creating awareness of the two groups' different points of view. After the immersive experience, the groups came together for a debrief session with neuroscientist Kris De Meyer, to allow players to find out what had been going on in the other room and to reflect on their decisions in that context. A post-debrief survey was conducted as a self-reported measure of perspective-taking.

Over the course of the experiment, participants in two different groups were regularly prompted to reflect on the perspectives and feelings of the different characters in the story, on which they had asymmetric information.

Profiles of the two main characters in the game. Each group followed the story from one character's perspective.



Nat
Jennings-Lee

Age:
16 since last week

Pronouns:
they/ them

Lives with:
Dad Stephen

Likes:
"Stories, animals, Oreos"

Worries about:
"Not knowing what I want to do with my life, being a bad vegan (I like ice-cream), my hair"

Good times with Jamie include:
"When he put peanut butter on a pepperoni pizza and then ate it all so he wouldn't have to admit it was disgusting!"

JAMIE
~~James~~
Stickson

Age:
16 just

Pronouns:
he/him

Lives with:
Mum Jenni and sister Ashleigh

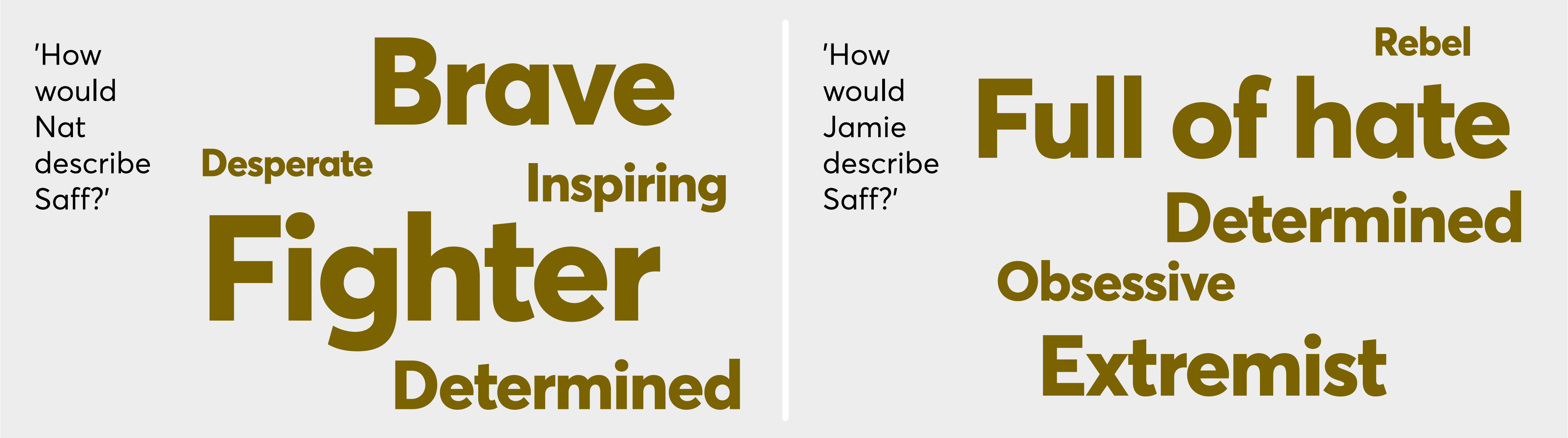
Likes:
"Rock climbing, cadets, Oreos"

Worries about:
"Money, Ashleigh (she's autistic and sometimes people stare at her), looking really young"

Good times with Nat include:
"When we looked after their aunt's dog (Shep) for a week, built an assault course for him in Nat's garden AND got paid for it!"

v. This concept is what is known as second-order theory of mind: it describes understanding what one person thinks about another person's thoughts.

Example of cognitive perspective-taking word selection by the two groups of participants: one followed the story from Nat's perspective and the other from Jamie's point of view. Saff is a third character about whom both rooms had different information.



'How would Nat describe Saff?'

Brave

Desperate

Inspiring

Fighter

Determined

'How would Jamie describe Saff?'

Rebel

Full of hate

Determined

Obsessive

Extremist

In all sessions run with students, the groups misidentified how the other character felt. However, when the session was run with adults, they largely got it right. This discrepancy suggests there is a case to be made for strengthening emotional perspective-taking skills among young adults, at a time when these skills may be in critical development.[18, 19, 20]

In the post-debrief survey, 84 per cent of the participants reported that the experiment made them think about taking other people's perspectives. For roughly half of the participants, taking part felt like watching a film or playing a game (implying a sense of doing something for fun and entertainment): an important consideration for engaging this age group. For about one-quarter of participants, it felt like making real-life decisions.

More research is needed to verify the impact and longevity of this approach on students' cognitive and emotional perspective-taking, and to further understand how their abilities differ from adults. However, these findings suggest that such novel methods are ripe for exploration by educationalists and cognitive scientists.

02

How to collaborate better

The complexity of social problems in a fast-moving world does not only require us to rethink the way in which we make decisions, but also how we can mobilise collective intelligence to collaborate more effectively.

Let the crowd self-organise

Experiment 5: Can a collective intelligence platform help orchestrate on-the-ground logistics in real time to reduce food waste?

Grantee: Hong Kong Baptist University, in collaboration with Foodlink Foundation.

Key finding: The introduction of Breadline led to a fourfold increase in volunteer efficiency and a 1.5-times increase in the food-donor base by streamlining processes for NGOs, donors and volunteers.

Who is this relevant for?

Anyone who needs to co-ordinate different actors in the context of time-critical or emergency situations, such as:

- Local governments.
- International organisations.
- Charities and NGOs.

Overview

In time-critical scenarios, effective co-ordination among different actors is key. Researchers at Hong Kong Baptist University wanted to test whether crowdsourcing could be extended beyond data to the movement of goods. The experiment found that decentralising co-ordination increased the efficiency of food rescue efforts fourfold.

Urban food rescue operations face the logistical challenge of collecting fluctuating volumes of supplies across disparate locations in a limited time.

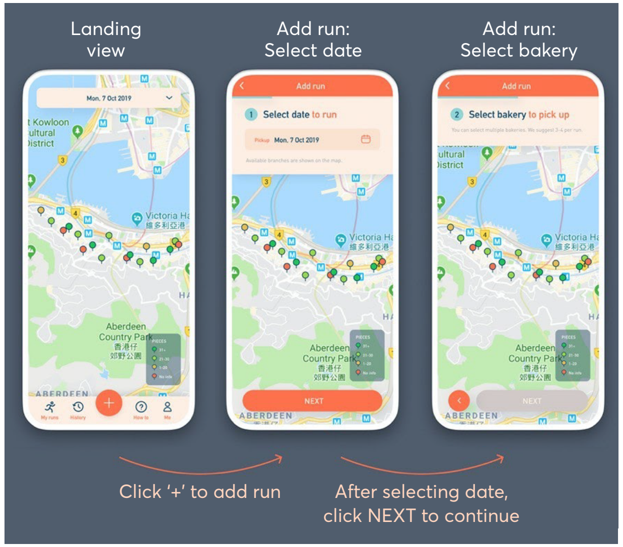
In this experiment, food rescue was used as a case study to explore the potential of networked intelligent actions – crowd-based logistics based on real time information. The goal was to test whether this would enable more effective mobilisation of resources compared to current centrally managed operations.

Hong Kong Baptist University created a digital platform called Breadline to address the following challenges:

1. **Spatial distribution** – The partner bakery chain had 240 bakeries located across the city.
2. **Temporal challenge** – 90 per cent of the bakeries close within 30 minutes of each other.
3. **Fluctuating availability** – It is uncertain how much bread there is left in each bakery until near closing (according to the NGOs, 20-30 per cent of all bakeries sell out).
4. **Intensive volunteer recruitment and management** – NGOs have to provide briefings, manually assign routes and assist volunteers during their run. The rate of volunteer retention is less than one per cent.

The Breadline platform allowed volunteers to choose their own routes and made previously undisclosed information available to them, such as the distribution of bakeries with food to be rescued and the amount of leftover bread each shop had in stock. Breadline also removed any shops from the list of available bakeries when volunteers indicated that they intended to pick up from a particular store.

The Breadline app as seen by volunteers



A comparison between the centrally managed process and Breadline

Existing centrally managed process	Using Breadline
Volunteers recruited by NGOs through outreach events	Volunteers recruited online through existing project networks
Volunteers assigned specific routes for bread collection by the NGO	Volunteers select their own route by choosing the bakeries to collect from directly on the platform
Volunteers get briefed at the NGO office on the day	No face-to-face briefing on the day
NGO staff provide support through WhatsApp messaging	No direct support by NGOs
Only NGOs can see the distribution of available bakeries	Both volunteers and NGOs can see the distribution of available bakeries
Information of leftover bread in each shop open to bakeries only	Information of leftover bread in each bakery accessible to volunteers
	Additional feature: volunteers can 'claim' shops which are then removed from the list of available bakeries

The experiment results suggest that shifting from a centralised to a decentralised model and sharing information with volunteers helped them to self-co-ordinate and make food rescue more efficient. On average, volunteers picked up **over four times as much** food to distribute when using Breadline compared to the standard process. The **increased transparency and access to new information** on Breadline enabled volunteers to co-ordinate better and work more efficiently in the following ways:

- **Avoiding the duplication of tasks**. On Breadline, volunteers were only shown bakeries that still had leftover bread, and bakeries were able to let users know how much bread they had available half an hour before closing. This way, journeys to shops where bread had been sold out or already picked up (so-called 'empty runs') could be almost completely avoided.
- **Reducing time-cost for volunteers**. Having access to more information allowed volunteers to integrate pickups into their daily commute or their lifestyle easily, which reduced the time-cost associated with the task. This could increase volunteer retention, something that will need to be studied over the long term.
- **Flexibly responding in real time**. Breadline also enabled volunteers to identify alternative routes and flexibly respond in real time. In the experiment, participants were diverging from original routes if more bread was available elsewhere, even if those bakeries were further away.

During this experiment, Hong Kong faced two major disruptions that impacted on the deployment and testing of the platform: the first was the protests sparked by the anti-extradition and anti-mask law; the second was the COVID-19 outbreak. Despite these challenges, Breadline was successfully deployed 12 times and resulted in 280 bakery pickups, saving 3,671 loaves of bread. A total of 87 volunteers were recruited to participate, and each volunteer picked up from 3.2 bakeries per run on average. Breadline was able to onboard idle bakeries (unclaimed by NGOs), increasing the donor base by 1.5 times without incurring more cost to Foodlink.

Volunteers in Hong Kong with leftover bread

The recent COVID-19 pandemic has demonstrated the importance of having platforms that can enable communities to respond to each other's needs. Although social media is able to broadcast information, it lacks the capacity to verify information and co-ordinate action, leading to duplication of efforts. The approach taken by Breadline deserves further exploration for its potential application to future community-led emergency and resilience initiatives.

Use semantic search to orchestrate intelligence sharing

Experiment 6: Can semantic search increase the efficiency of human rights defenders in building a shared database on digital rights?

Grantee: HURIDOCS, in collaboration with Social Media Exchange (SMEX).

Key finding: FT_HR (fasttext) was the most accurate and fastest semantic search algorithm of the five tested by the team.

Who is this relevant for?

- Human rights defenders.
- Legal professionals.
- Intergovernmental organisations.
- Governments.
- Anyone working with large document collections or databases and wanting to collaborate.

Overview

To do their work well, human rights professionals depend on quickly and effectively filtering relevant information from large amounts of often unstructured data. Document collections are incredibly diverse in terms of language and sets of terminology, which makes this process even more difficult and makes collaboration between organisations challenging. HURIDOCS, a human rights organisation, tested whether semantic search algorithms can help to orchestrate intelligence sharing and enable human rights defenders to work together more effectively.

Making informed decisions in today's saturated world requires not only the ability to access information, but also the ability to filter it based on its relevance to your needs. This is true across many different sectors, including human rights. From legislation and legal cases to progress reports and diplomatic commitments, there exists an abundance of human rights information that can potentially support efforts to protect people's fundamental dignity and freedoms. For this information to be meaningful, however, human rights defenders need to be able to find it quickly and efficiently – and human rights defenders are a diverse group, working across multiple languages and sets of terminology, with varied professional expertise. Consequently, joining different databases – and therefore accessing the knowledge within them – is challenging and can hinder collaboration between different groups who are all pursuing the same goal.

Machine learning-based **semantic search** is a powerful tool that could mitigate this challenge. Unlike searches that only match exact keywords, this search technique seeks to give users better results by understanding the concept that they are looking for in order to show related terms. For example, if a user were to search ‘education’, words like ‘schools’ and ‘universities’ would also appear.

This experiment set out to test what would happen if machine learning-based semantic searches were applied to a database of human rights information, and to understand which approach would deliver the best results for users. Would it increase the efficiency and accuracy of tagging documents in the database?

The first part of the experiment compared different search algorithms by assessing their accuracy and computation time on different human rights databases. The second part studied the impact of a machine learning-based semantic search tool on five users attempting to curate a collection of documents around the specific human rights issue of digital rights.

In the first part of the experiment, the researchers found that one algorithm, FT_HR fasttext, outperformed the others in keyword searches. USE, an algorithm trained on large collections of documents, performed best when it came to searching similar sentences. The sample size for the second part of the experiment was too small to draw meaningful conclusions, and the experiment contained too many uncontrolled variables, such as the differing professional roles of the participants and levels of familiarity with the Uwazi database. Uwazi is the platform that was used in the experiment for creating and organising collections of human rights information.

Further research is needed to determine whether semantic search – a technique that is widely applied across for-profit sectors – could render human rights information more accessible and inclusive, blurring the traditional divides between silos of language and expertise in a constructive and collaborative way.

What are natural language processing (NLP) and semantic search algorithms?

Everything humans express carries huge amounts of explicit and implicit information: through the context, our selection of words or the tone we use. NLP is defined as the ability of a computer to understand and translate human-generated text and potentially simulate language.[2] Most NLP techniques rely on machine learning to derive meaning from human languages.

Semantic search uses NLP. It addresses the problem that normal search indexes only come up with results based on the content of the information that has been indexed but do not show results that are similar in meaning. Semantic search algorithms are trained on large databases to extract meaning from sentences or paragraphs and can therefore provide more detailed and comprehensive search results. In its experiment, HURIDOCS tested five different semantic search algorithms:

- **FT_Wiki**: a fasttext model pretrained on Wikipedia documents.
- **FT_HR**: a fasttext model trained on human rights legal data.
- **USE**: Universal Sentence Encoder.
- **USE_large**: Universal Sentence Encoder large.
- **BERT**: currently the state-of-the-art natural language model.

Quality was measured by accuracy, precision, recall and the F1 score. The assumption was that language models trained on large collections of documents, such as BERT or USE, would yield more accurate and reliable results than methods based solely on word embeddings.

Use AI to connect similar ideas and like-minded people

Experiment 7: Can using NLP help more citizens get their views heard on digital democracy platforms?

Grantee: Alan Turing Institute for Data Science and AI, in collaboration with the University of Warwick and Consul.

Key finding: Using NLP reduced the time it took to find similar citizen-generated proposals by 40.9 per cent or 58.7 per cent, depending on the method used.

Who is this relevant for?

- Cities and governments using the Consul digital platform and other digital democracy initiatives.

Overview

Consul is one of the most successful open-source digital participation platforms, used by over 100 cities and organisations all over the world. While engagement rates are high, very few citizen proposals reach the public support necessary to be considered by the local government. The Alan Turing Institute tested whether using NLP techniques could help to connect similar proposals and like-minded users on Consul. The goal was to increase collaboration between citizens on the platform and thereby increase the number of proposals passing the required threshold of support. The experiment found that NLP-enabled features reduce the time it takes to find similar proposals.

As trust in traditional democratic institutions declines, digital democracy platforms such as Consul offer a way to build new relationships between citizens and policymakers. The growing popularity of such platforms is a welcome development, but the large volume of citizen-generated content is becoming increasingly difficult to navigate. Ideas are often spread over dozens of different proposals from citizens, leading to duplication of effort.

As a consequence, very few citizen proposals have so far reached the levels of support necessary to be taken into account by the local government. This experiment tested whether using NLP techniques could help to connect similar proposals and users on Consul, to make it easier for like-minded citizens to collaborate towards common goals.

Originally, the experiment was planned to take place on the Consul platform used by the city of Madrid. Following a change in the ruling political party, though – which saw Madrid's Department of Citizen Participation, Transparency and Open Government downgraded – the team had to pivot to a lab-based experiment.

The data used for the experiment was nevertheless retrieved from public Consul data sets of Madrid City Council and included proposals, comments on proposals and manually entered proposal tags by users. Based on this data set, and using different NLP methods, the researchers developed four new features to help users navigate the platform more effectively: automatically generated **tags**, a list of **related proposals**, a **summary of the comments** and a **related users** list.

The experiment found that using NLP could **improve the effectiveness of citizen participation** on digital platforms, mainly by reducing the time it takes to find similar proposals.

- Using automatically generated **tags** reduced the time to find similar proposals by 40.9 per cent compared to using tags manually entered by citizens.[vi] However, if the author of a proposal had used very defined and descriptive tags, this enabled participants to find similar proposals more quickly because the narrow labels allowed people to search a much smaller set of proposals.[vii]
- The **list of related proposals** generated through NLP seemed to be a useful feature because it made finding similar proposals both easier and significantly faster (58.7 per cent) compared to the original Consul version without such lists. Overall, both features also seem to increase the quality of the results: the proposals people found on the enhanced platform through tags and lists of related proposals were considered more similar by participants.
- Despite being a closely related task, finding similar users from the **related user list** was considered more difficult by experiment participants and took longer than searching for similar proposals using the related proposals list.
- The **comment summaries** were not considered very relevant or useful by the participants, and the quality between summaries varied highly, indicating a need for improvement and further testing of this feature.

Features based on NLP techniques can be a useful complement to purely human participation. They enable people to find relevant proposals faster and help to connect citizens with similar interests and a shared vision. To understand whether those features indeed lead to higher levels of collaboration and more proposals achieving success, the enhanced version of Consul will need to be tested in practice.

The team has secured continuation funding for the project from the Warwick–Turing postdoctoral associate research scheme, and they are in conversation with several local authorities about live-testing the new features on Consul.

vi. On the original version of Consul, citizens can use their own words and terms when tagging or drafting proposals and leaving comments, so the variety of terms and language used makes it more difficult to identify similar proposals that already exist.

vii. However, authors of proposals rarely use very specific tags: analysis of manually entered tags in the data set showed that the 50 most used tags account for 80 per cent of all tags used.

How was this tested?

To develop and apply new features to the Consul platform, the researchers first had to understand the variety of topics covered by the proposals on Consul. By applying **topic modelling** to the proposals in the data set, they generated 40 new topics, with each topic defined by the five most relevant terms used in the proposals. For example, Topic #5 was represented by (underground line bus lift stop). Each proposal in the data set could be represented as a combination of different topics, which allowed the researchers to identify the proposals that best represented each topic. The three most representative proposals for Topic #5 were 'New underground line in the east of Madrid', 'Public bicycles in Almudena underground' and 'Underground in Valderrivas neighbourhood'.

This way, the algorithm could also identify which five terms out of the 200 were most representative for each proposal – these were made the **proposal tags**.

Because the topic modelling uniquely defined each proposal in the topic space, their relative similarity to each other could be calculated, generating a **list of most similar proposals** for each proposal. The researchers then applied text summarisation algorithms to identify the key sentences across the user comments, generating a **comment summary** for each proposal. Finally, the team generated a **related users list** for each user, based on all proposals and comments created by that user.

To evaluate the effectiveness of the new features compared to the original platform, the experiment participants were randomly assigned to either the treatment or control condition and asked to fulfil the following tasks:

1. Look for one or more proposals they would like to support: to do this, they chose one or more **tags** in successive steps, until they found four proposals that were similar.
2. Find proposals that are similar to the one they have already found: to do this, they used the **related proposals lists** until they found four proposals that were similar.
3. Decide which proposal they would like to support by reading the **summary of the comments** others had left.
4. Find like-minded citizens: to do this, they used the **related users lists** until they found four users that were similar.

The effectiveness of each feature was measured by different variables, such as how long subjects took to complete a task, the quality of the result and the perceived difficulty of the task by the participant. After the experiments, participants were asked to fill in a questionnaire to measure subjective satisfaction with the enhanced Consul version.

03

How to make better use of citizen-generated data

Crowd insights and data harnessed through collective intelligence have led to breakthroughs in traditionally elite professional fields, such as science research and healthcare. It has changed the way that laws get made and enabled us to improve our understanding of situations in real time. But as the volume of data increases, so do the challenges of navigating and analysing it.

Make it easier to process citizen data

Experiment 8: Can natural language understanding help encourage quicker political response to citizen views on digital democracy platforms?

Grantee: CitizenLab.

Key finding: Improving the manual processing of citizen input on the CitizenLab platform increased content moderation efficiency by 36 per cent and increased the percentage of citizens receiving feedback on their ideas from 35 per cent to 84 per cent.

Who is this relevant for?

- Local governments.
- National policymakers.
- Digital democracy platforms.

Overview

This experiment aimed to improve policymakers' uptake of citizen-generated ideas and insights from digital democracy platforms. It tested the impact of using natural language understanding (NLU) to automate the analysis of unstructured citizen inputs (ideas, arguments and votes) on the CitizenLab platform. It found that new manually controlled solutions helped to significantly increase the number of citizen ideas receiving feedback.

What is NLU and how is it different from NLP?

NLP is defined as the ability of a computer to understand and translate human-generated text and potentially simulate language.[2] It is an umbrella term to explain the whole process of turning unstructured data into structured data.

NLU is a subset of NLP. NLU focuses primarily on machine-reading comprehension by interpreting the meaning that the user communicates. It allows computers to understand and interpret human language.

Many cities operate digital democracy platforms to collect ideas and opinions from citizens. These large collections of unstructured data need to be analysed, which is difficult and time-consuming for city administrators. This can often discourage the use of such data in decision-making processes. This experiment tested whether using NLU would make it quicker for public administrators to analyse insights from digital democracy platforms, and therefore increase their responsiveness to citizen ideas.

Based on interviews with city administrators in Belgium, Denmark and the Netherlands, CitizenLab made several improvements to its platform, which allowed it to perform tasks differently.

How was this tested?

In-depth interviews with different cities revealed that they found the most time-consuming part of using the CitizenLab digital democracy platform to be following up and processing citizen input. To address these issues, CitizenLab developed a mix of manual and automated solutions.

Manual solutions:

- **Moderation view**: Maintaining a full overview of citizen input, including comments, was identified as the most urgent requirement by cities. In response, CitizenLab improved the manual moderation by adding a dedicated back-office page to easily follow up on all new citizen input and take action if required.
- **Feedback flow and counter functionality**: This is a new feature that allows city officials to assign ideas to civil servants. It also reminds the civil servants which citizen input still requires feedback.
- **Mass-feedback flow with personalisation**: This enables administrators to group ideas that require similar feedback and provide generic feedback copy that allows for some personalisation (e.g. adding the name of the citizen).

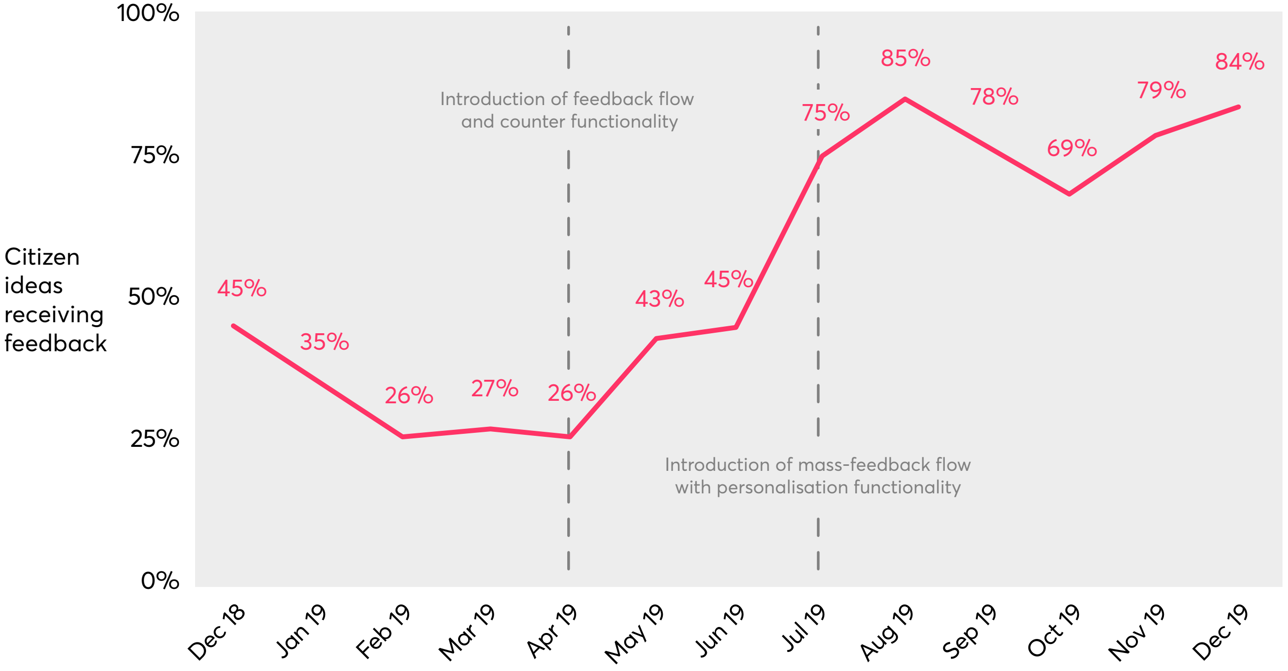
NLU-assisted automated solutions:

- **NLU-assisted moderation**, where potentially harmful or irrelevant citizen input is auto-flagged.
- **NLU-assisted idea classification**, where the AI automatically detects topics that are related to a specific piece of citizen input.
- **NLU-assisted clustering feature**, which automatically groups similar content in an unsupervised way.

All those solutions were validated and tested through prototypes or mock-ups with multiple cities. The aim of the validation was to understand which of the solutions increased the efficiency of the process and whether local authorities perceived them as helping to solve their key problems.

Improving the platform with **manual solutions** helped to substantially reduce the time that city officials spent on moderating and processing citizen input. The efficiency of comment moderation increased by 36 per cent. As this task had been reported as the most time-consuming one by city officials, this change constitutes a significant improvement. Assigning ideas to civil servants and creating customisable feedback templates increased the percentage of citizen ideas receiving feedback from 35 per cent before the features were introduced to 84 per cent after implementation (as shown in Figure 5).

Figure 5: Percentage of citizen ideas receiving feedback from local authorities between December 2018 and December 2019



The experiment did not succeed in fully testing the NLU-assisted solutions within the time frame of the reporting period. However, insights from the seven in-depth interviews with city administrators found that the NLU-assisted clustering feature highlighted trends and needs of which cities weren't aware. For example, two cities working on new strategies reported that they already had a structure and underlying objectives for their plan in mind when processing the citizen input. This meant that, when they processed the input manually, there was a strong tendency for the officials to flex every citizen idea to fit into the existing structure and objectives. Using the automated clustering functionality, however, prevented this confirmation bias and revealed trends and needs that had not been identified previously. This is an area deserving of more research.

The interviews also revealed that most city administrators preferred to have more efficient manual solutions over automated ones, and that it was important for them to feel 'in control' of the process of responding to citizens.

Following this experiment, CitizenLab secured a €500,000 grant from the Brussels Region. This will allow it to further explore the potential of NLU to transform citizen input into actionable policy recommendations and increase its uptake by policymakers.

Look for ways of unlocking new insights from existing data sets

Experiment 9: Is it possible to identify the most effective feedback from data gathered from an online maths assessment platform? And can behavioural prompts encourage teachers to adopt 'best practice'?

Grantee: Behavioural Insights Team and Eedi.

Key finding: By analysing a large volume of feedback from teachers and data on student performance, it was possible to identify particularly effective feedback.

Who is this relevant for?

- Ed tech companies.
- Educationalists.

Overview

Teachers give feedback to students all the time, but there is little evidence about what works. This experiment aimed to find out whether it is possible to uncover the type of feedback from teachers that leads to the greatest improvement in students' performance in maths by analysing data from an online assessment platform. It also tested whether behavioural prompts could encourage teachers to adopt this way of giving feedback.

There is little evidence of which marking and feedback practices work best, leading teachers to waste valuable time. Online education platforms hold large amounts of data from both students and teachers. The homework platform Eedi, for example, currently stores over 54,000 pieces of individual feedback that teachers have given to their students. Such online tools help teachers to assess their pupils' homework easily, but don't yet allow them to tap into the collective wisdom of their colleagues to increase student performance.

This experiment took 14,649 pieces of feedback generated by teachers and analysed the improvement of students who received it based on their performance across two consecutive quizzes. From this, the team was able to extract a small number of pieces of feedback that had been particularly effective on specific maths questions.

In the second part of the experiment, teachers were encouraged to use this specific feedback through a prompt on the website, displayed next to relevant questions. An email drawing on the behavioural concept of procedural transparency – explaining how the effective feedback had been identified – was also tested.

There was no statistically significant difference in the amount of feedback teachers left when they received a behaviourally informed email compared to the control group, who received an email without the behavioural concept. Overall, the feedback feature on the Eedi website had much lower usage than anticipated; just two per cent of teachers used it. Consequently, the team collected insufficient data to draw any meaningful results.

Use machine learning to analyse, catalogue and preserve eyewitness footage

Experiment 10: Can machine learning turn crowdsourced footage of airstrikes into legal evidence of British weapons being used in Yemen?

Grantee: Hillary Rodham Clinton School of Law, Swansea University, in collaboration with Global Legal Action Network (GLAN), VFRAME, Syrian Archive and Yemeni Archive.

Key finding: Using a machine learning algorithm trained on synthetic data is helping to identify British-manufactured cluster munitions in citizen-generated footage of airstrikes in Yemen more quickly.

Who is this relevant for?

- Human rights defenders and legal practitioners.
- Journalists.
- Civil society organisations.

Overview

Citizen-generated video footage from eyewitnesses is increasingly being shared online and used to discover mass human rights violations. However, the sheer volume of data and the challenge of verification makes it difficult for human rights investigators to use this evidence in court cases. This experiment developed a machine learning algorithm trained on synthetic data to support investigators to analyse and identify relevant footage. It also developed a secure way to store and manage evidence collections. It is hoped that this experiment will lead to the first case to show that citizen-generated evidence can be admitted before the UK courts.

Journalists, the police and human rights investigators increasingly rely on crowdsourced digital evidence. In human rights investigations, footage gathered by eyewitnesses and shared online – often in real time – can provide valuable insights into the occurrence of human rights violations. However, manually reviewing the data is time-consuming and expensive, and can cause trauma to investigators due to the disturbing nature of the content of the footage. In addition, the willingness of civilian witnesses or organisations to share this data depends on the recipients' ability to handle sensitive information securely. For courts to afford the information weight, it is also crucial to demonstrate that the evidence hasn't been tampered with.

In this experiment, researchers developed and tested a machine learning tool trained on synthetic data to analyse citizen-generated footage videos of war crimes in Yemen. The tool was trained to identify a specific type of cluster munition, which has been used by Saudi forces in the Yemen war and therefore could be used as supporting evidence in legal proceedings. The researchers hypothesised that the machine learning tool could identify cluster munitions in videos more efficiently than manual filtering.

What is synthetic training data and why is it special?

As highly specific objects, such as cluster munitions, do not appear in enough videos to **train** an algorithm, the researchers had to generate synthetic image training data. They developed a 3D model that matched the target images, based on information about the texture, colour and dimensions of the weapon from a handful of photos found online. Once the model was complete, they designed environments that match the landscapes in the real-world target videos and simulated and randomised lighting and camera positions to create a training data set.

The lack of a real-world data set on this type of cluster munition also makes it difficult to **validate** the performance of the algorithm trained on synthetic data. The researchers are now experimenting with the possibility of using staged photographs of 3D-printed munition copies to create a validation dataset. If successful, this constitutes a new paradigm for training and validating using only synthetic data with virtually no access to natural training data. But, as with all computer vision algorithms, object detection is a probabilistic determination and, to confirm whether each video actually shows evidence of illegal cluster munitions, manual oversight is still required. The algorithm is therefore not replacing the researcher; they work in tandem.

The team also developed a secure means to manage their collection of citizen-generated evidence of human rights violations in Yemen by integrating Uwazi, open-source software for building and sharing the human rights document collection with the Digital Evidence Vault. The Digital Evidence Vault is software that simplifies the preservation of digital content from websites like YouTube and social media, and timestamps each piece of evidence that is being saved. This step is particularly crucial because images and videos containing possible evidence are often quickly removed, either by the platform or the poster.

The experiment is part of an ongoing investigation. Results to date suggest that machine learning tools can be successfully developed and applied to support humans in data analysis, even where there is limited ground truth data, by using synthetic data sets for training. Early tests with the machine learning tool have shown that it can process 5,750 videos per day. This means it would take an estimated 8-10 days to analyse the Yemeni Archive for documentation of a specific cluster munition. For comparison, a human analyst would need to watch each video, which would take an estimated 400 days.[viii]

However, expert verification of videos flagged by the algorithm as containing cluster munitions is still necessary, and the above numbers don't take into account the time needed for this process. In addition, the object detector is still undergoing benchmarking for accuracy, so while using a machine learning tool to identify potential evidence saves time, the precision of the algorithm is still unclear.

Although we are still awaiting the final conclusions from this experiment, it is a compelling example of crowd–machine collaboration. Using the methods developed in this experiment, GLAN submitted a detailed dossier of evidence to the UK Government in August 2019.[21] The evidence was a combination of 'closed' evidence compiled by Mwatana, a Yemeni independent human rights group, and open-source evidence. The dossier showed the extent of unlawful attacks carried out by the Saudi/UAE-led coalition in Yemen and argued for the UK Government to cease all existing British government licences to export arms to Saudi Arabia for use in Yemen.[22] Integrating Uwazi with the Digital Evidence Vault allowed GLAN and Mwatana to collaborate in compiling and analysing the evidence.

viii. This calculation assumes a data set size of 400,000 videos at an average of 1.5 minutes and 25 frames per second. Prior to running the object detection algorithm, each video is pre-processed using scene analysis, which removes visually similar frames that are redundant. The actual detector algorithm can analyse approximately 50–60 images per second. Therefore, after pre-processing, analysing each video for cluster munitions only takes 2 seconds per video.

04

How to increase effectiveness of participation in collective intelligence initiatives

The success of collective intelligence initiatives using methods such as crowdsourcing and citizen science is directly linked to their ability to engage crowds of people taking part in collecting and analysing data. When collective intelligence initiatives rely on dedicated volunteers or contributions over longer periods of time, it is particularly crucial to sustain engagement and keep dropout levels low.

Don't assume you already know how to get the best out of your crowdworkers

Experiment 11: What improves the performance of crowdworkers analysing tweets for disaster relief and recovery efforts?

Grantee: University of Southampton.

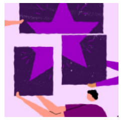
Key finding: Higher pay for crowdworkers did not always result in more or higher-quality work, and even had an adverse impact on labelling accuracy. In addition, crowdworkers favoured repetition over variation in tasks and were more likely to respond to feedback from other crowdworkers than experts.

Who is this relevant for?

- Citizen science projects and (paid) crowdsourcing platforms.
- International and humanitarian organisations.

Overview

In humanitarian emergencies, organisations increasingly rely on crowdsourcing data analysis from volunteers. However, volunteer engagement is often short-lived. Focusing on crowdworker analysis of text- and image-based tweets sent during hurricanes Harvey, Irma and Maria, researchers at the University of Southampton assessed the impact of different tactics on volunteer motivation and performance. The experiment found that higher pay did not improve the quality of work performed, and that crowdworkers were more likely to respond to feedback received from other crowdworkers than an expert.



Thanks to new data sources, such as social media posts and drone footage, aid agencies and local authorities have access to more data than ever before when assessing and co-ordinating their responses. Crowdsourcing is a proven method for collecting and analysing data quickly. For example, it took volunteers only a few hours to map 4,000 disaster-related events after the 2010 Haiti earthquake.[23] The longer a crisis lasts, however, the more difficult it gets to mobilise a critical mass of contributions. This is partly because many people are attracted to volunteering when media coverage is high, which is usually directly after disaster strikes.

In this experiment, the University of Southampton explored the impact of **different payment levels, varying the type of tasks** and **varying the difficulty of tasks** and **receiving feedback** on task accuracy and crowdworker motivation.

While many crowdsourcing tasks are performed by volunteers, paid crowdworkers are increasingly playing an important role in analysing information during disasters. While there will be lessons for working with volunteers, this experiment focused on paid crowdworkers recruited through Amazon Mechanical Turk.[ix]

The experiment was originally designed to test how to maintain crowdworker engagement in the long term. As it proved difficult to engage a sufficiently large crowd of participants during the experiment, the researchers were forced to change their focus from tactics to sustain long-term engagement to short-term incentives to increase crowd accuracy and motivation.

The researchers developed four experiments based on publicly available data sets composed of both text- and image-based tweets sent during hurricanes Harvey, Irma and Maria. Experiment participants were asked to analyse the content and relevance of those tweets for disaster relief efforts.

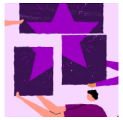
What is the role of crowdwork in analysing and labelling data?

From content moderation to translation, data entry and classification, crowdworkers work on a range of tasks and are often needed when computational processes reach their limits.[24] In content moderation on social media for example, where decisions about the appropriateness of a post or tweet are nuanced, humans are indispensable: if there is no clear 'yes' or 'no' answer, it is currently impossible to train an algorithm with sufficient accuracy to automate this type of task.

The huge role crowdworkers play in labelling large data sets makes them fundamental to the latest advances in AI. Machine learning algorithms often depend on someone curating the training data and putting it into a structured format, so that a computational process can learn from it.[x] Yet, often these people remain the untold story behind the success of AI.[2] In addition, there has been substantial criticism of this type of work and the platforms and companies that reinforce current structures, particularly with regard to the lack of regulations around wages, working conditions and social security, such as sick pay.

[ix] Amazon Mechanical Turk allows requesters to design and implement crowdsourcing tasks – known as HITs (Human Intelligence Tasks) – and recruit participants from Amazon's pool of crowdworkers. Mechanical Turk was selected because, at the time of the experiment, it had the highest average pay of paid crowdsourcing platforms.

[x] This process is called supervised machine learning.



How was this tested?

The researchers developed four experiments, based on text- and image-based tweets related to hurricanes Harvey, Irma and Maria.

- **Experiment 1 (E1)** – Participants were asked to complete 10 daily tasks (1 unpaid recruitment task and 9 paid follow-up tasks). Each task consisted of a set of five text- and/or image-based tweets where participants were asked to **identify names of people, locations and the names of occupations**. Workers were then randomly assigned to one of three payment conditions but were unaware of the specific amounts offered and to which condition they were assigned.[xi]
- **Experiment 2 (E2)** – Workers were first asked to **classify sets of text-based tweets** according to content from a list of options (e.g. request for help, weather update) and perceived difficulty. Tasks were then clustered to form a simple and difficult variant of the task. The performance in those two variants was compared to a control group, where the tasks were presented at random. Unlike in E1, workers were all paid the same amount ($0.45) per task.
- **Experiment 3 (E3)** – Like E2, with the difference that workers were asked to **classify sets of image-based** tweets according to content from a list with a greater number of options (e.g. hazards, people, repairs in progress), which made the task more difficult and complex than E2. Workers were paid $0.20 per image classified and could classify as many images as they wished.
- **Experiment 4 (E4)** – Like E3, but workers were prompted with **feedback** after they submitted their answers. One group was shown existing responses from other workers and the other was shown responses from a domain expert. Workers were asked if they wished to adjust their responses based on this feedback and, if not, whether they felt the feedback was incorrect or matched their existing answer.

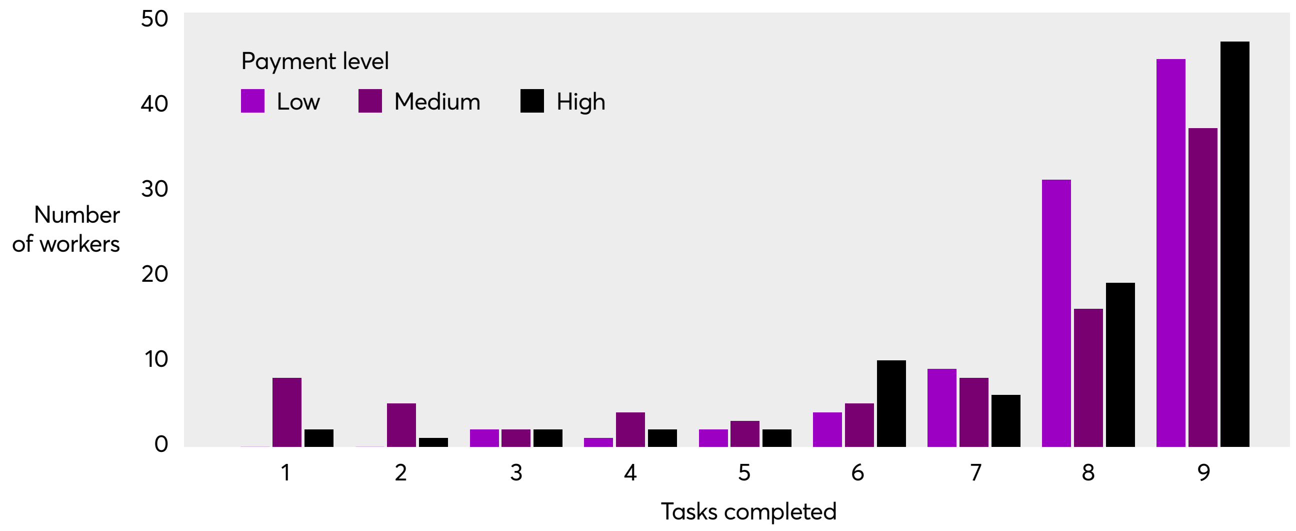
The researchers found that:

- **Higher pay did not always result in more or higher-quality work**. To understand the impact of different pay levels on crowdworker performance, participants in the first experiment received different amounts of money for their work. Workers that were paid a below-average amount did as much work, or even more work, than participants that were paid above-average or average amounts (Figure 6). This may be because people had to do more work to earn the same money, or that the primary motivation was altruism and money acted as a disincentive. Furthermore, the results indicate that higher payments have a potential **negative** impact on the quality of crowdwork. Workers who were paid less were significantly **less** likely to make clearly false, unfinished or low-effort submissions than those who were paid above average.
- **Task difficulty affected accuracy but not levels of participation**. The second and third experiments gave participants tasks with different levels of difficulty. Results found that, while task difficulty did not impact participation levels, it affected the accuracy of submissions. Unsurprisingly, people who were assigned easy tasks performed better than those who were assigned difficult tasks when classifying text-based tweets. When classifying image-based tweets, the group receiving tweets with random difficulty performed better than those receiving difficult tweets. The researchers didn't find any difference between the easy and difficult or easy and random tasks.

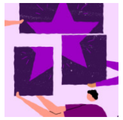
[xi] Those assigned to the low condition were offered $0.80 for each of the nine follow-up tasks completed, equivalent to approximately $4.80 as an hourly wage and around the lower end of the average of worker earnings in Mechanical Turk. Those in the medium condition were offered $1.05, or an equivalent of $6.50 as an hourly wage, at the higher end of the hourly average for Mechanical Turk. Finally, those assigned to the high condition received $1.30, equivalent to an hourly wage of $7.80, an above-average amount for Mechanical Turk and above the US federal minimum wage.

- **Worker accuracy tended to be low across all levels of difficulty**, with many tasks receiving correct responses from one worker only or no correct responses at all. When analysing images, workers particularly seemed to overlook details or misinterpret elements. One potential reason for this inaccuracy was that the high level of detail in many of the images made the tasks more difficult than expected. This resulted in low submission accuracy, even for images that participants perceived as easy. This finding suggests that crowdworkers may require support to identify features in disaster management tasks, and that they should be asked to identify only a small number of specific, clear features.
- **Crowdworkers favoured repetition over variation in tasks**. Workers preferred easily repeatable tasks that could be achieved rapidly. In the follow-up survey conducted by the researchers, workers reported being more motivated by the lack of task variation because the simplicity of the task and growing familiarity with it allowed them to improve and learn much quicker than when the task was more varied.
- **Crowdworkers are more likely to change their assessment based on feedback from other crowdworkers than from experts**. The fourth experiment tested what feedback would make crowdworkers change their initial assessment. It found that they were more likely to be influenced by feedback on their assessment from peer crowdworkers than feedback from experts. When workers were provided with existing answers from other crowdworkers, 41 per cent of workers adjusted their answer to match that of the crowd, but only 26 per cent chose to do so when shown feedback from an expert.

Figure 6: Number of tasks completed by workers in different payment conditions (E1)



The purpose and underlying incentive structures of collective intelligence initiatives that engage volunteers differ from paid crowdwork. Collective intelligence projects typically articulate a collective purpose, whereas paid crowd-labelling work focuses on individual gains by paying for each completed microtask.[2] These differences should be taken into account before transferring the insights from this experiment to a different context.



Use AI for smarter matching of people to projects

Experiment 12: Can AI personalise project recommendations to citizen scientists to increase their engagement?

Grantee: University of Edinburgh, in collaboration with SciStarter.

Key finding: The algorithm that suggested the least-known projects increased user engagement on SciStarter the most.

Who is this relevant for?

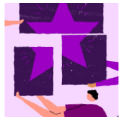
- Citizen science platforms and any collective intelligence initiative that wants to encourage participants to participate in a variety of tasks or projects.

Overview

Citizen science platforms host thousands of projects to which volunteers can contribute. Faced with a large number of options, it is difficult for individuals to find the project that best suits their preferences and skills. This challenge can affect their motivation and satisfaction. Researchers at the University of Edinburgh tested different recommendation algorithms on the citizen science platform SciStarter to see which would increase user participation in projects. The experiment found that an algorithm that used a technique called matrix factorisation increased user engagement on the platform the most by recommending less well-known projects.

Citizen science has been remarkably successful in advancing scientific knowledge. It has helped to speed up data analysis, increased sample collections and sometimes even led to historic scientific breakthroughs.[25, 26] Citizen science platforms offer thousands of different projects, covering everything from astronomy to marine debris and linguistics. For citizen scientists, identifying the projects that match their skills and interest is difficult due to the wide variety of projects that many platforms offer.

Researchers at the University of Edinburgh collaborated with the citizen science platform SciStarter to test four different recommendation algorithms in order to match users with projects that best suit their interests and capabilities. SciStarter has over 65,000 registered users and offers 3,000 projects on its platform.



How was this tested?

The team developed four different recommendation algorithms, three of which were based on collaborative filtering, one of the main approaches for recommendation systems.[xii] Collaborative filtering is a method that ranks the relevance of a project to a user based on their past activities (e.g. the projects they have searched for, clicked on or participated in previously) and by comparing those to the actions of other users. All algorithms were trained and tested for accuracy with historical data from SciStarter.

- **Algorithm 1**, called **item-based collaborative filtering**, is based on project similarity. It ranks a project more highly for User A if the project is similar to other projects that they have contributed to in the past.
- **Algorithm 2**, called **user-based collaborative filtering**, is based on user similarity. Two users, A and B, are considered similar if they have contributed to the same projects in the past. It is assumed that, if this is the case, A is likely to share B's interest in a new project. This algorithm therefore ranks a project as more relevant for User A if similar users to them (such as User B) have also contributed to the project.
- **Algorithm 3**, called **matrix factorisation** (more specifically, a technique known as singular value decomposition or SVD), directly computes the relevance of a new project to a target user by modelling the user. It predicts the user's rating for each project, and projects with the highest predicted ratings are recommended.[xiii] For example, the project Asteroid Mappers (an online project designed to identify craters on the asteroid Vesta) is recommended to User A because this project has the highest predicted rating for this target user out of all the projects. This is mainly because in the past User A has participated in the projects Moon Mappers (an online project designed to perform science tasks on Mars) and StallCatchers (an online project designed to accelerate Alzheimer's research), which are similar to the Asteroid Mappers project. Also, Asteroid Mappers best describes the correlation of User A with other users.
- **Algorithm 4** is not based on collaborative filtering. Here, all projects are sorted by their **popularity**, which is simply measured by the number of users who contribute to each project.

The performance of the algorithms was evaluated using different measures:

- **Precision** of the algorithm – the proportion of recommended projects that the user eventually interacted with (either by clicking on them and/or participating) during the experiment time frame, with 100 per cent being that the users interacted with all recommended projects at some point.
- **Recall** – the proportion of projects with which the users interacted that were also recommended, with 100 per cent signifying that all the projects with which the users interacted were recommended by the algorithm (which means the user didn't interact with any non-recommended projects).

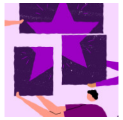
To better understand the recommendation quality, additional metrics were used. Two metrics were used to measure the **number of interactions with recommended projects**:

- **Click and act** – the number of users clicking on a recommended project and interacting with it later.
- **No click but act** – the number of user actions where the user did not click on a recommended project but did interact with it at a later stage.

The researchers also measured the **number of interactions with non-recommended projects ('Act but no rec')**. Finally, the experiment also measured the users' engagement with the recommendation tool, such as clicking on the project, on a project's image or on the 'next' and 'previous' buttons (**'tool interaction'**).

xii. The other approach that is generally used for recommendation systems is the content-based method, which bases recommendations on shared items and user characteristics (such as location, targeted age group or project task).

xiii. This algorithm is different from the others because it's not just based on defining a similarity measure. SVD uses a matrix where the users are rows, projects are columns and the entries are ratings that represent the relevance of the projects to the users. It estimates the relevance of a target project for a user by maintaining user and project models that include hidden variables that can affect how users choose items. These variables have no semantics, they are simply numbers in a matrix. In reality, aspects like gender, culture and age affect the relevancy a lot, but the researchers did not have access to this data. The user-project matrix is obtained by multiplying the user model and the item model matrices.



In the experiment, the algorithms were deployed on the platform for 41 days. A total of 125 SciStarter users were randomly assigned to five cohorts. Four cohorts were provided with recommendations based on one of the algorithms. The fifth cohort functioned as the **control group**, which did not get any recommendations, but was provided with the SciStarter's promoted projects, which appear on the website as a standard feature.

Three out of four algorithms performed as well or better than the control group when measuring the interaction of users with the recommended projects. This indicates that users are more engaged with SciStarter and the projects when receiving personalised recommendations. From the four algorithms tested, the matrix factorisation algorithm, which predicted users' rating for each project and suggested projects with the highest predicted ratings, performed best.[xiv] This means it increased the activity of volunteers on the platform the most. According to a follow-up survey, users were generally more interested in projects they had never heard about. Based on this result, the researchers suspect that the reason for the higher impact of the matrix factorisation algorithm was that it introduced less popular projects that were generally less well known to users than the ones recommended through the three other algorithms in the experiment.

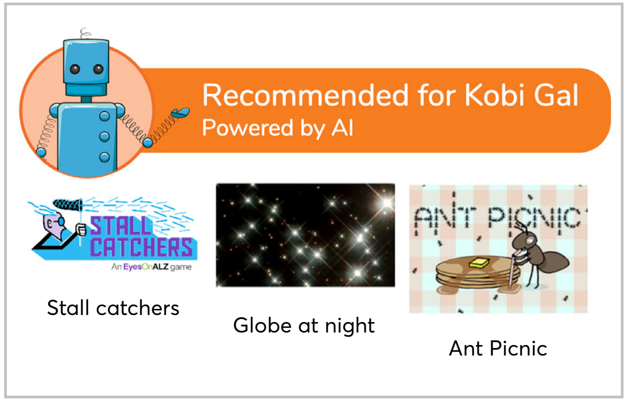
However, the precision of all four algorithms was low: the proportion of recommended projects that the user eventually interacted with (either by clicking on them and/or participating) was below 12 per cent for all of them. This shows that, while AI-powered recommendations can help better match users with projects, there remains room for improvement.

The most successful algorithm is now deployed on the SciStarter platform as a permanent feature and user participation in projects increased during the experiment period.

Results for the five experiment cohorts. Each cohort consisted of 25 users

Cohort (25 users each)	Click and act	No click but act	Act but no rec	Tool Interaction	Precision	Recall
A1 (item-based collaborative filtering)	13	3	56	25	2.9%	8.8%
A2 (user-based collaborative filtering)	36	6	37	72	11.3%	28.2%
A3 (matrix factorisation)	54	8	64	124	11.4%	24.4%
A4 (popularity)	38	1	29	99	9.1%	22.5%
Control group	15	1	29	29	9.3%	24.2%

Screenshot of recommended citizen science projects as suggested by the recommendation algorithms



[xiv] This finding is consistent with existing literature, which considers matrix factorisation the leading algorithm in the domain of recommendation systems.[27, 28]

Conclusion

The diversity of the experiments illustrates the breadth of the concept of collective intelligence and its diverse applications. Importantly, these experiments also demonstrate the cutting edge of crowd–machine collaboration. Some of the insights from the experiments can be directly applied; others pave the way for further research in the field.

Nesta's Collective Intelligence Grants Programme created space for new ideas to be tried and tested, and sometimes to fail – a crucial way of accelerating learning in the field. It enabled non-profit organisations, universities and companies to develop new and unique approaches and collaborate in new ways.

Despite its potential, collective intelligence design is still a nascent area for research funding and is dwarfed by investments in AI. There are, however, signs of growing interest in the field of collective intelligence. In 2019, for example, the European Commission launched a new €6 million fund[29] for applied research in this area.

Last year, Nesta also announced a second round of funding through the Collective Intelligence Grants Programme. For this round, we are delighted to co-fund 16 experiments with a total fund pot of £500,000, in partnership with Cloudera Foundation, Omidyar Network and Wellcome.

However, the fact remains that there are currently no large-scale funding opportunities in the UK for collective intelligence research and development. This gap could be filled by UKRI and foundations, as well as by public funders integrating collective intelligence into existing AI funding programmes.

The first major funder to put £10 million into this field has the opportunity to make a lasting impact on it. The COVID-19 pandemic has demonstrated the many uses of collective intelligence: from symptom-tracking apps to open-source production of medical equipment. It is an idea whose time has come.

More information on the experiments

For more information about the experiments, please reach out to the grantees directly.

Experiment 1: Will autonomous agents help people co-operate better?

Contact: Prof Tom Lenaerts (Tom.Lenaerts@vub.be)

The researchers have also published a research paper on their experiment: arxiv.org/abs/2003.07317

Experiment 2: Will Swarm AI help politically polarised groups come to more collectively acceptable decisions than traditional voting methods?

Contact: David Baltaxe (david@unanimous.ai)

Experiment 3: Can multi-agent AI systems help us make better decisions by balancing out our biases?

Contact: Vito Trianni (vito.trianni@istc.cnr.it)

Experiment 4: Can immersive digital storytelling help improve the ability of young people to see other perspectives?

Contact: Rachel Briscoe (rachel@fastfamiliar.com)

Fast Familiar published a detailed overview of the experiment on its website: cdn.fastfamiliar.com/pdf/200423%20IIWY%20info.pdf

Experiment 5: Can a collective intelligence platform help orchestrate on-the-ground logistics in real time to reduce food waste?

Contact: Daisy Tam (daisytam@hkbu.edu.hk)

Hong Kong FoodWorks provides more information about the experiment and the food system in Hong Kong more broadly: hkfoodworks.com

Experiment 6: Can semantic search increase the efficiency of human rights defenders in building a shared database on digital rights?

Contact: info@huridocs.org

Experiment 7: Can using NLP help more citizens get their views heard on digital democracy platforms?

Contact: Prof Rob Procter (Rob.Procter@warwick.ac.uk)

Experiment 8: Can natural language understanding help encourage quicker political response to citizen views on digital democracy platforms?

Contact: Karel Verhaeghe (karel@citizenlab.co)

CitizenLab published several blogs on its website throughout the experiment: www.citizenlab.co/blog/civic-engagement/the-human-factor-helping-governments-tap-into-collective-intelligence

www.citizenlab.co/blog/product-update/natural-language-processing-at-citizenlab-how-machine-learning-can-transform-citizen-engagement-projects

Experiment 9: Is it possible to identify the most effective feedback from data gathered from an online maths assessment platform? And can behavioural prompts encourage teachers to adopt 'best practice'?

Contact: Lal Chadeesingh (lal.chadeesingh@bi.team)

Experiment 10: Can machine learning turn crowdsourced footage of airstrikes into legal evidence of British weapons being used in Yemen?

Contact: Prof Yvonne McDermott Rees (yvonne.mcdermottrees@swansea.ac.uk)

The Global Legal Action Network (GLAN) published a detailed blog on the experiment: www.glanlaw.org/airstrike-evidence-database-yemen

Experiment 11: What improves the performance of crowdworkers analysing tweets for disaster relief and recovery efforts?

Contact: Prof Elena Simperl (elena.simperl@kcl.ac.uk)

Experiment 12: Can AI personalise project recommendations to citizen scientists to increase their engagement?

Contact: Prof Kobi Gal (kgal@inf.ed.ac.uk)

The SciStarter podcast hosts the episode 'Smart Recommendations in SciStarter', with more details and background on the experiment: scistarter.org/podcast

Bibliography

Bang, D. and Frith, C. (2017). 'Making better decisions in groups.' Royal Society Open Science, 4(8): 170193–22. https://doi.org/10.1098/rsos.170193

Berditchevskaia, A. and Baeck, P. (2020). 'The Future of Minds and Machines: how artificial intelligence can enhance collective intelligence.' Centre for Collective Intelligence Design, Nesta, London. https://media.nesta.org.uk/documents/FINAL_The_future_of_minds_and_machines.pdf

Bösser, T. (2001). 'Autonomous Agents. International Encyclopedia of the Social & Behavioral Sciences.' 1002-1006. https://doi.org/10.1016/B0-08-043076-7/00534-9

Cooper, D.J. and Kagel, J.H. (2009). 'Other-Regarding Preferences: A selective Survey of experimental Results.' Handb. Exp. Econ. 2.

Corea, F. (2019). 'Distributed Artificial Intelligence (Part II): A Primer On MAS, ABM And Swarm Intelligence.' https://www.forbes.com/sites/cognitiveworld/2019/03/21/distributed-artificial-intelligence-part-ii-a-primer-on-mas-abm-and-swarm-intelligence/#7e8493cb261f. Accessed April 2020.

Crone, E.A. and Fuligni, A.J. (2020). 'Self and Others in Adolescence.' Annual Review of Psychology, 71. https://doi.org/10.1146/annurev-psych-010419-050937

De Melo, C., Marsella, S. and Gratch, J. (2018). 'Social decisions and fairness change when people's interests are represented by autonomous agents.' Auton. Agent. Multi. Agent. Syst. 32, 163–187. https://doi.org/10.1007/s10458-017-9376-6

De Melo, C., Marsella, S. and Gratch, J. (2019). 'Human Co-operation When Acting Through Autonomous Machines. Proceedings of the National Academy of Sciences of the United States of America.' https://doi.org/10.1073/pnas.1817656116

Dodell-Feder, D., & Tamir, D. I. (2018). 'Fiction reading has a small positive impact on social cognition: A meta-analysis.' Journal of Experimental Psychology: General, 147(11): 1713–1727. http://dx.doi.org/10.1037/xge0000395

Gower, S. (2014). Netflix prize and svd. http://buzzard.pugetsound.edu/courses/2014spring/420projects/math420-UPS-spring-2014-gower-netflix-SVD.pdf

Gray, M. L., and Suri, S. (2019). 'Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.' Houghton Mifflin Harcourt.

Hamman, J.R., Loewenstein, G., and Weber, R.A. (2010). 'Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship.' Am. Econ. Rev. 100, 1826–1846.

Hunt, L. (2007). 'Inventing Human Rights: a History.' W.W. Norton.

Koriat, A. (2015). 'When two heads are better than one and when they can be worse: The amplification hypothesis.' Journal of Experimental Psychology: General, 144(5): 934–950. http://dx.doi.org/10.1037/xge0000092

Lin, T., Ha, S.Y., Li, W. et al. (2019). 'Effects of collaborative small-group discussions on early adolescents' social reasoning.' Reading and Writing 32: 2223–2249. https://doi.org/10.1007/s11145-019-09946-7

MacDonald, E.A., Donovan, E., Nishimura, Y. et al. (2018). 'New science in plain sight: Citizen scientists lead to the discovery of optical structure in the upper atmosphere.' Science Advances, 4 (3). https://doi.org/10.1126/sciadv.aaq0030

Milinski, Manfred et al. (2008). 'The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change. Proceedings of the National Academy of Sciences of the United States of America.' 105(7): 2291–94. https://doi.org/10.1073/pnas.0709546105

Peach, K., Berditchevskaia, A., and Bass, T. (2019). 'The Collective Intelligence Design Playbook. Centre for Collective Intelligence Design.' Nesta, London. https://media.nesta.org.uk/documents/Nesta_Playbook_001_Web.pdf

Sadek, R. A. (2012). 'Svd based image processing applications: state of the art, contributions and research challenges.' International Journal of Advanced Computer Science and Applications, 3 (7). https://arxiv.org/ftp/arxiv/papers/1211/1211.7102.pdf

Schwenck, C., Göhle, B., Hauf, J. et al. (2014). 'Cognitive and emotional empathy in typically developing children: The influence of age, gender, and intelligence.' European Journal of Developmental Psychology, 11:1, 63-76. https://doi.org/10.1080/17405629.2013.808994

Silke, C., Brady, B., Boylan, C. and Dolan, P. (2018). 'Factors influencing the development of empathy and pro-social behaviour among adolescents: A systematic review.' Children and Youth Services Review, 94, 421-436. https://doi.org/10.1016/j.childyouth.2018.07.027

Tavoni, A. et al. (2011). 'Inequality, communication, and the avoidance of disastrous climate change in a public goods game.' Proc. Natl. Acad. Sci. 108: 1–10. https://doi.org/10.1073/ pnas.1102493108

Trouille, L., Lintott, C.J., and Fortson, L.F. (2019). 'Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems.' PNAS 116 (6) 1902-1909. https://doi.org/10.1073/pnas.1807190116

Willcox, G., Rosenberg, L., Burgman, M., and Marcoci, A. (forthcoming). 'Prioritizing Policy Objectives in Polarized Societies using Artificial Swarm Intelligence.' In the Proceedings of the 2020 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA).

Wintour, P. (2020). UK receives report documenting Saudi cover-up of unlawful Yemen airstrikes. https://www.theguardian.com/world/2019/aug/15/report-documenting-saudi- cover-up-of-unlawful-airstrikes-in-yemen-submitted-to-uk. Accessed June 2020.

Wolf, M. (2018). 'Reader, come home: the reading brain in a digital world.' Harper.

Yang, D., Zhang, D., Frank, K., et al. (2014). 'Providing real-time assistance in disaster relief by leveraging crowdsourcing power.' Personal and Ubiquitous Computing, 18(8), 2025-2034. https://doi.org/10.1007/s00779-014-0758-3

Endnotes

[1] Peach, K., Berditchevskaia, A., and Bass, T. (2019).

[2] Berditchevskaia, A. and Baeck, P. (2020).

[3] Milinski, Manfred et al. (2008).

[4] Tavoni, A. et al. (2011).

[5] De Melo, C., Marsella, S. and Gratch, J. (2018).

[6] Bösser, T. (2001).

[7] Hamman, J.R., Loewenstein, G., and Weber, R.A. (2010).

[8] Cooper, D.J. and Kagel, J.H. (2009).

[9] De Melo, C., Marsella, S. and Gratch, J. (2019).

[10] Willcox, G., Rosenberg, L., Burgman, M., and Marcoci, A. (forthcoming).

[11] Bang, D. and Frith, C. (2017).

[12] Koriat, A. (2015).

[13] Corea, F. (2019).

[14] Hunt, L. (2007).

[15] Wolf, M. (2018).

[16] Dodell-Feder, D., & Tamir, D. I. (2018).

[17] Lin, T., Ha, S.Y., Li, W. et al. (2019).

[18] Crone, E.A. and Fuligni, A.J. (2020).

[19] Silke, C., Brady, B., Boylan, C. and Dolan, P. (2018).

[20] Schwenck, C., Göhle, B., Hauf, J. et al. (2014).

[21] https://mwatana.org/en/mwatana-glan-submission-to-uk/ Accessed June 2020.

[22] Wintour, P. (2020).

[23] Yang, D., Zhang, D., Frank, K., et al. (2014).

[24] Gray, M. L., and Suri, S. (2019).

[25] Trouille, L., Lintott, C.J., and Fortson, L.F. (2019).

[26] MacDonald, E.A., Donovan, E., Nishimura, Y. et al. (2018).

[27] Sadek, R. A. (2012).

[28] Gower, S. (2014).

[29] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/ict-54-2020 Accessed June 2020.

nesta

58 Victoria Embankment
London EC4Y 0DS

+44 (0)20 7438 2500

information@nesta.org.uk

@nesta_uk

www.facebook.com/nesta.uk

www.nesta.org.uk