Images in this presentation generated using ChatGPT unless otherwise stated

EDRC
Energy Demand Research Centre

Keeping people in the loop
How (not?) to co-design agents for energy flexibility

9 September 2026

User-Centre Energy Systems Academy webinar

Michael Fell

Principal Research Fellow, University College London

edrc.ac.uk

UK Research and Innovation

Economic and Social Research Council

Engineering and Physical Sciences Research Council

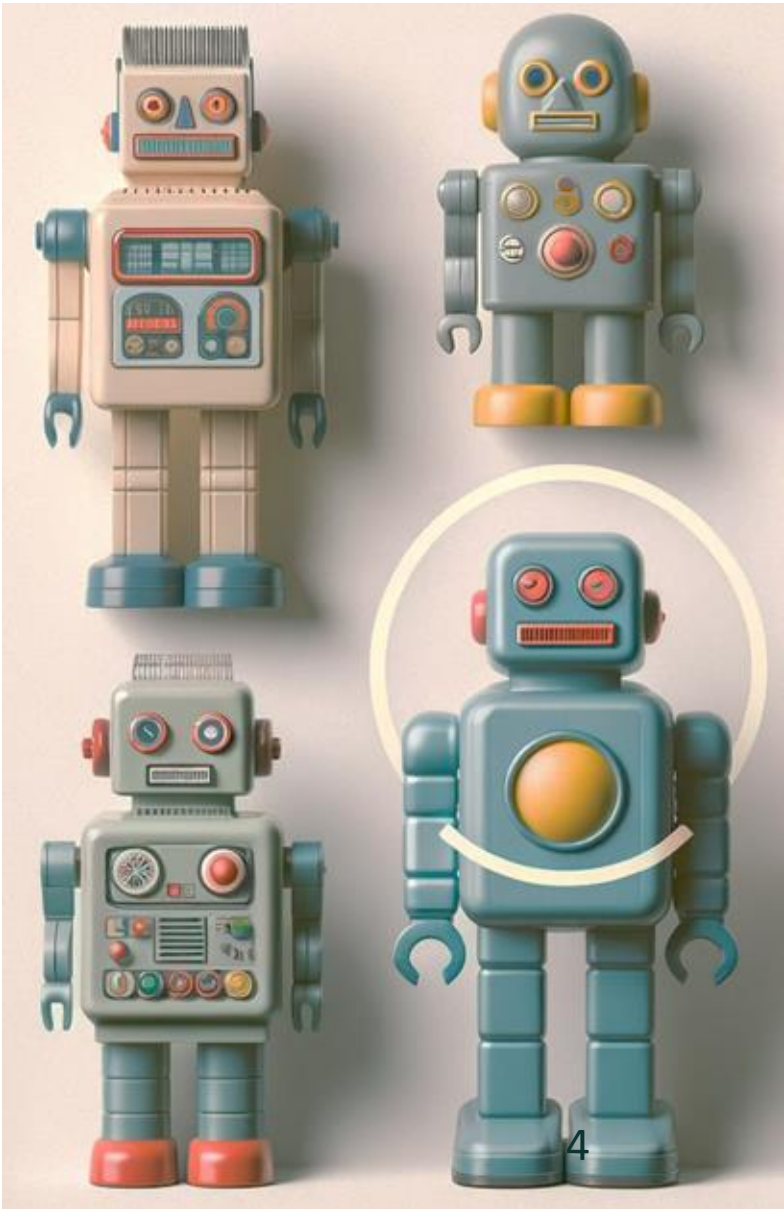
UCL

To cover

- Background on use of LLM-based agents to simulate individuals
- My work co-designing AI preference agents – what I did and found, with interpretation and implications

Background

- Large Language Models (LLMs) can be prompted to simulate human responses
- Argyle et al. 2023 introduce the “silicon sampling” approach, whereby multiple LLM agents are endowed with a selection of characteristics, e.g. republican, female, then respond to question prompts
- Find that GPT-3 can replicate viewpoints of demographically varied populations
- Lots of subsequent research across domains…



I used a basic Python program to endow an LLM agent with a range of demographic, attitudinal and personality characteristics, e.g.:

You are a UK householder completing a survey in a hurry, so rely more on fast, intuitive judgement. You are male, aged 70, with a medium household income, 3 people in your household, you highest level of education is 'further', you are at home most of time, and you are a social tenant. You have these attitudes: high environmental concern, low risk aversion, medium social trust, left of centre politics, high place attachment, high economic rationality, and your innovation adoption status is 'late majority'. Your personality has traits of low extraversion, medium agreeableness, high conscientiousness, high neuroticism, and medium openness to new experience.

…then respond to a question from an existing study:

A new energy offer is available which allows you to buy energy directly from homes and businesses with their own solar panels. Energy bought this way is slightly cheaper than what you get from your usual supplier. If you participate, you could meet **around a quarter** of your household's electricity needs through the offer. You would buy electricity directly from homes and businesses **located on the streets in your neighbourhood**. You would continue to buy the rest of your energy from your current supplier. Would you sign up to participate in this offer if it was available to you today?

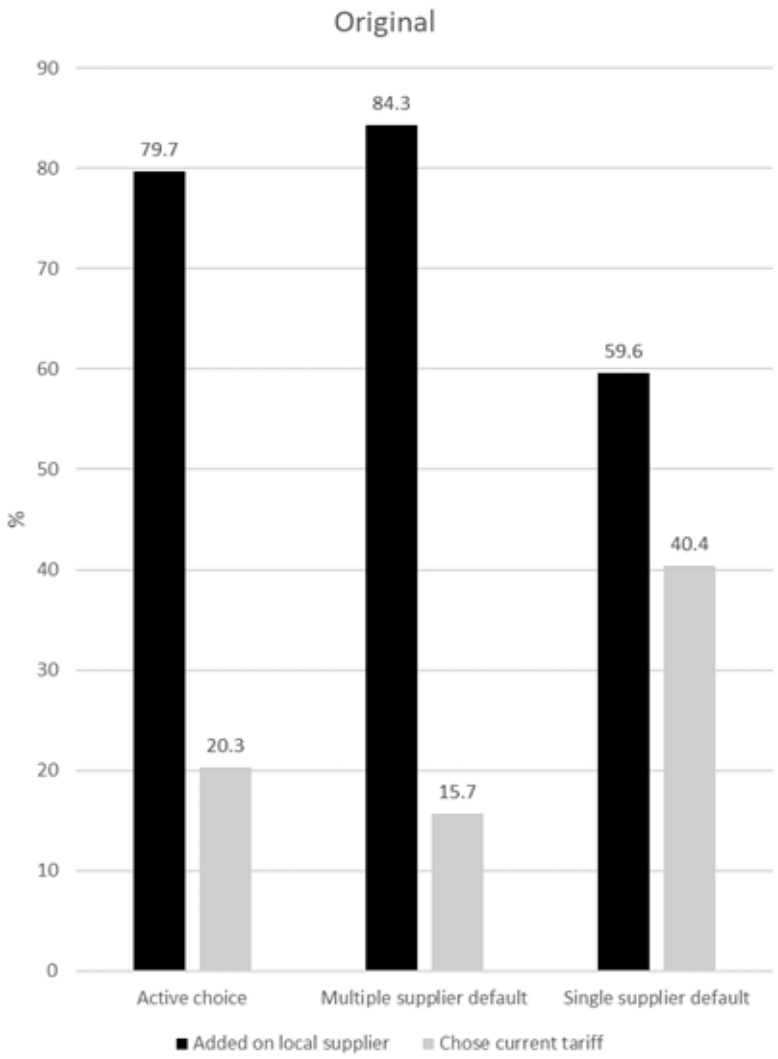
…with specific instructions about how to respond:

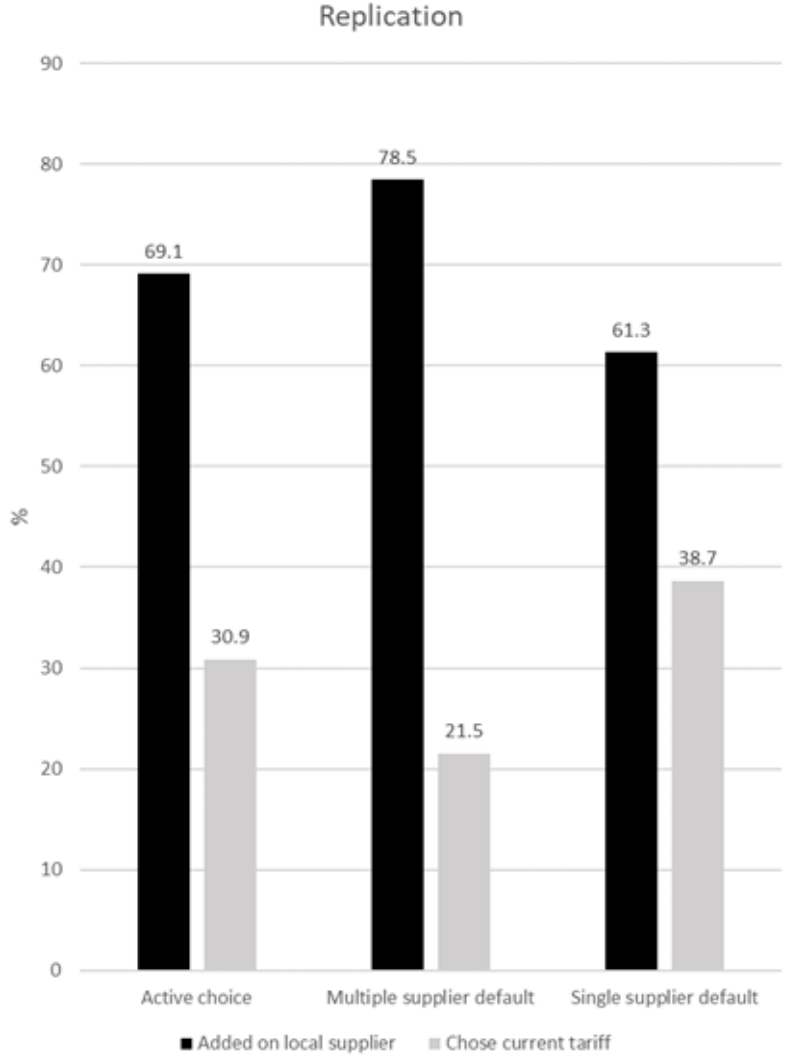
Respond in JSON format. The first JSON object should be a short (<50 word) explanation of your reasoning, called 'explanation', drawing on your demographic, attitudinal, and personality characteristics. Then, output your decision on whether or not you would participate, in a JSON object called 'decision', with response options yes=1 or no=0 (return an integer). Your decision must be consistent with your explanation. Example output as follows: 'explanation': 'explanation text here', 'decision': integer

Then a new agent with a new set of characteristics is generated, a response received, and so on – providing a set of responses from a diverse sample.

Response speed	Decision to participate	Explanation
At leisure	Yes	As a homeowner with medium environmental concern and a general willingness to adopt new innovations, I would be open to the idea of buying energy directly from local homes and businesses with solar panels. I have a medium place attachment, so supporting local energy production aligns with my values.
	No	As a homeowner with medium risk aversion and low environmental concern, I prefer to stick with my current supplier for reliability and stability. Despite the cost savings, I value the consistency of my current energy supply over the potential benefits of a cheaper, decentralized option.
In a hurry	Yes	I prefer to support local energy initiatives and as an early adopter, I am open to trying new energy solutions. The slight cost saving and the community aspect appeal to me.
	No	I don't have high environmental concern, and I prefer to stick with what I know. Plus, I'm not usually at home to engage with my neighbours in this way.

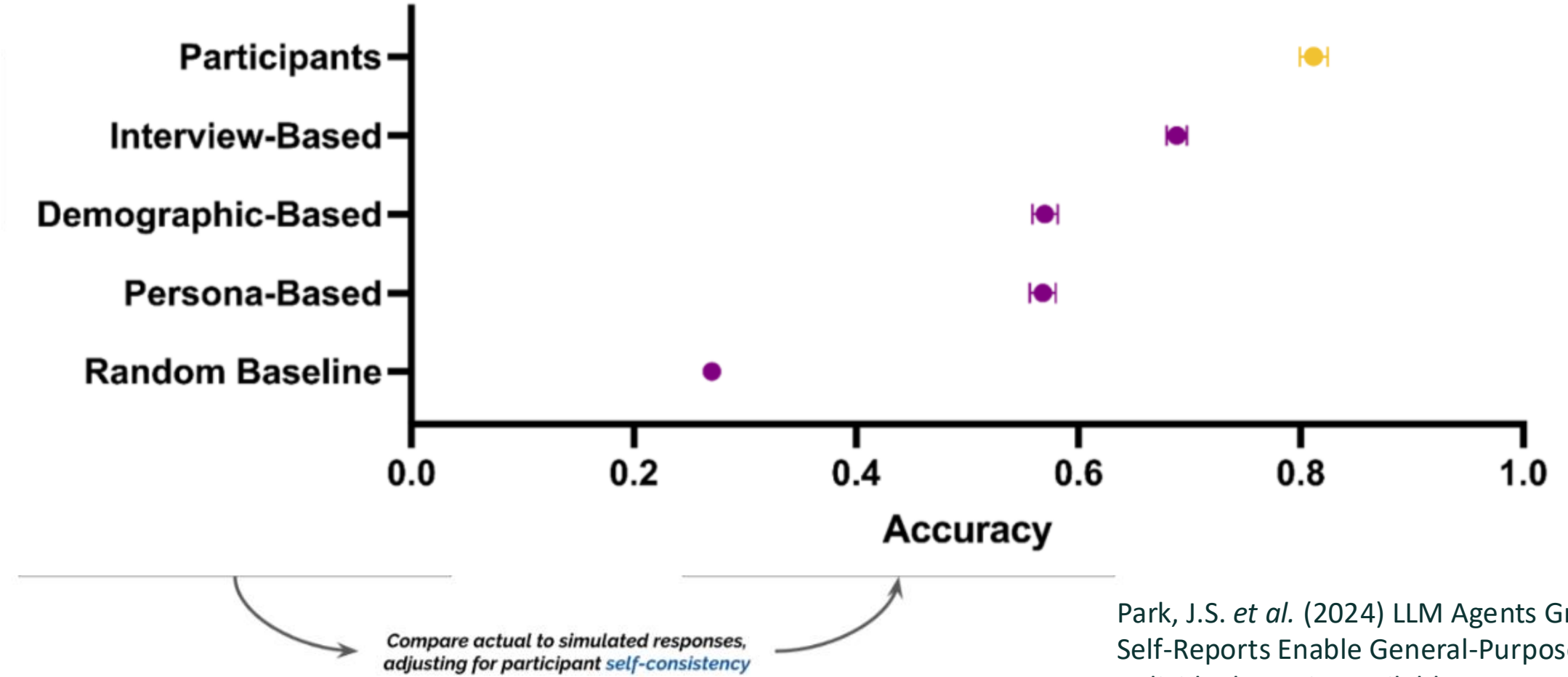
Replicating Watson et al., 2020





- "Higher fidelity", all but one stat significant differences replicate
- Including reproduction on an unexpected result (overturning of default effect)

General Social Survey



Park, J.S. *et al.* (2024) LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. arXiv. Available at: https://doi.org/10.48550/arXiv.2411.10109.

EDRC Energy Demand Research Centre

cheaper, faster, more “inclusive”

- simpler – supporting survey design, or synthetic data
- more ambitious – user research, agent-based modelling, action

Syntheticusers Uxia deepsona

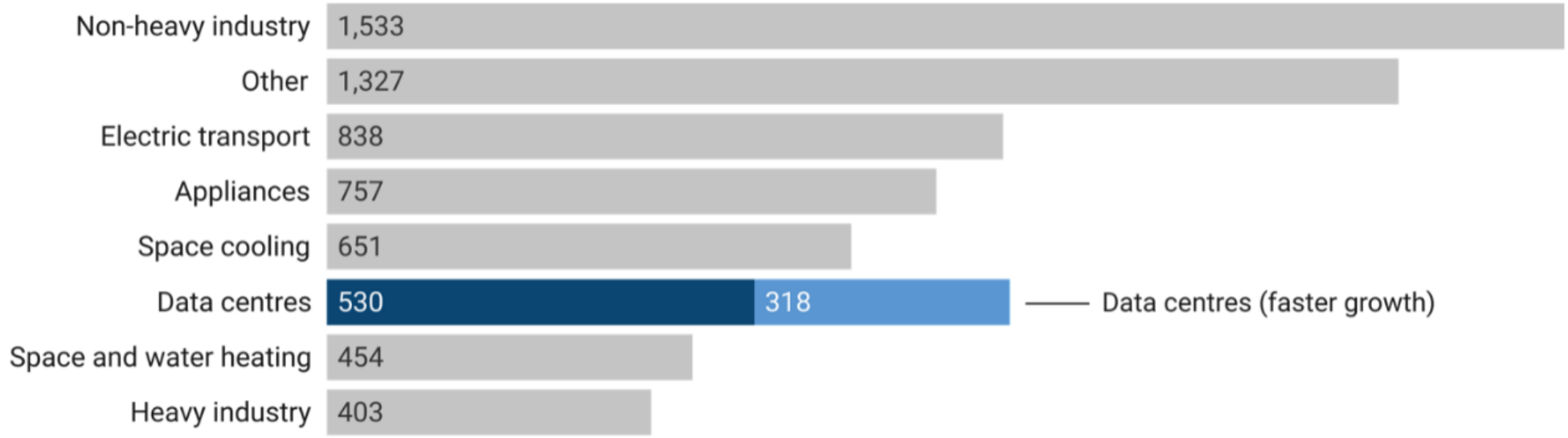
Key challenges

- Validity – especially where limited comparator data
- Misrepresentation and stereotyping
- Excluding real voices and grounded experiences
- Uncritical/malicious use

Energy use

Around a tenth of the growth in global electricity demand is expected to come from data centres over the next five years

IEA projections of electricity demand growth between 2024 and 2030, TWh

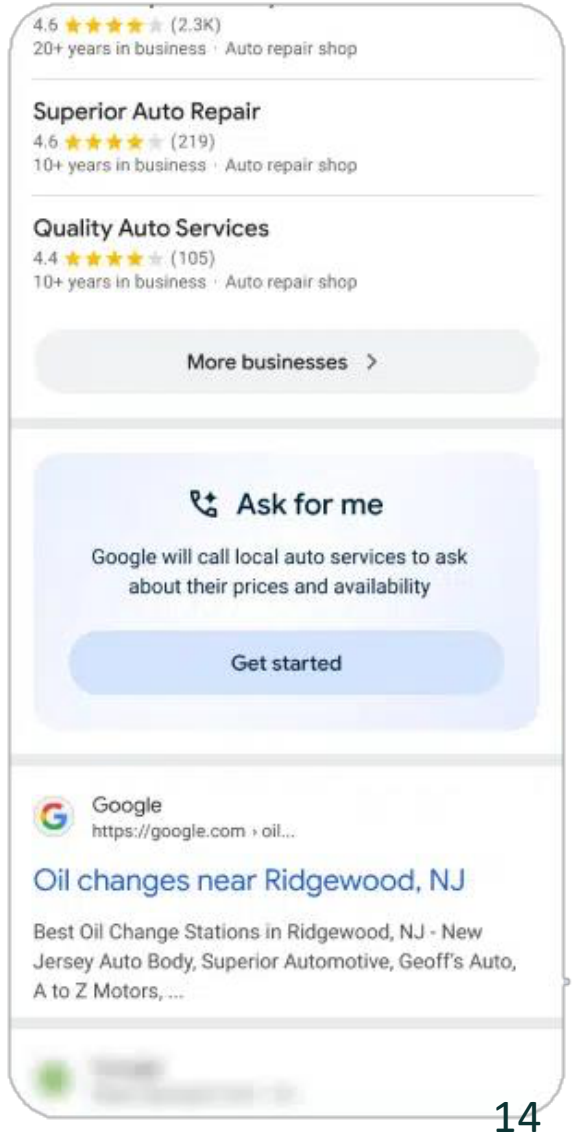


Source: IEA energy and AI report.

Electricity demand growth in IEA scenarios, in TWh between 2024 and 2030, due to data-centre expansion (dark blue) – including a scenario in which expansion happens faster (light blue) – as well as from other sectors (grey). Source: IEA energy and AI report.

Things to come

- Humans will increasingly be represented by AI agents
- Delegation of decision-making is nothing new, but scope and extent of a different order – for example…



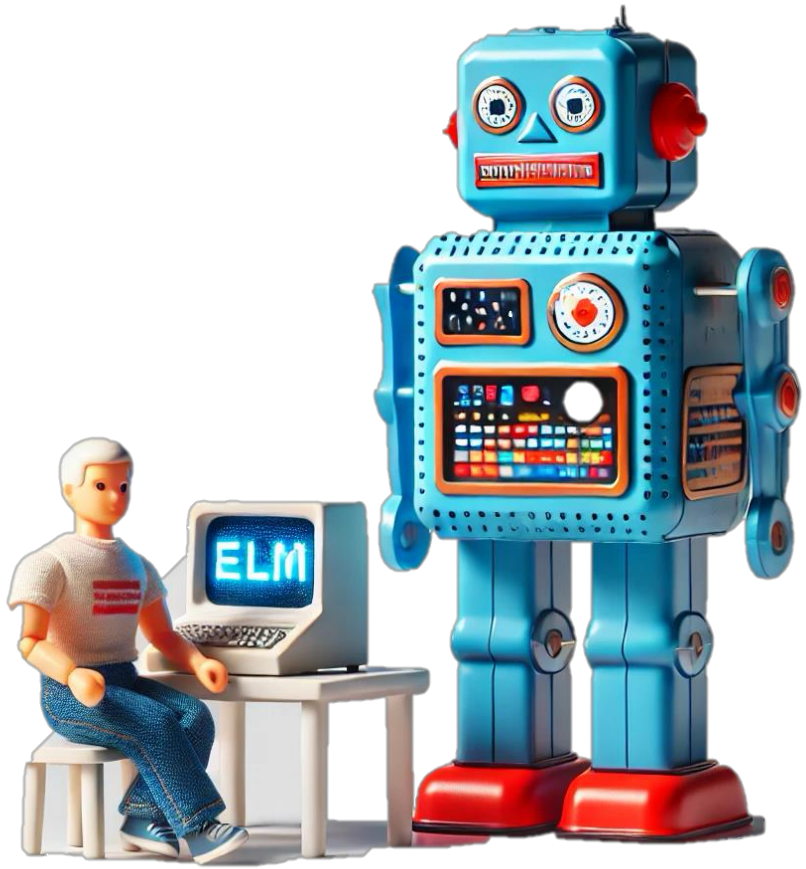
SHARED
NEIGHBOURHOOD
BATTERY

Increasingly agent behaviour = human behaviour

Understanding processes by which agent behaviour is developed and overseen is therefore critical

Co-designing LLM agents

- I conducted research working with people to co-design agents to represent them
- Potential benefits include: better objective and perceived validity, better inclusivity
- Key questions:
  - How do people engage in and perceive an LLM preference agent co-design process?
  - How do people perceive the performance of their co-designed agents?
  - How effective are co-designed agents at accurately representing people's preferences?
- Potential risks around: data privacy, misuse of agents…



Codesigning preference agents

“[that is] exactly the thought process I went through”

— P12

“summarising my 50/50 grey area view, it doesn’t reflect it very well”

— P11

“[The agent] literally captured everything in a much more succinct way than I probably explained here”

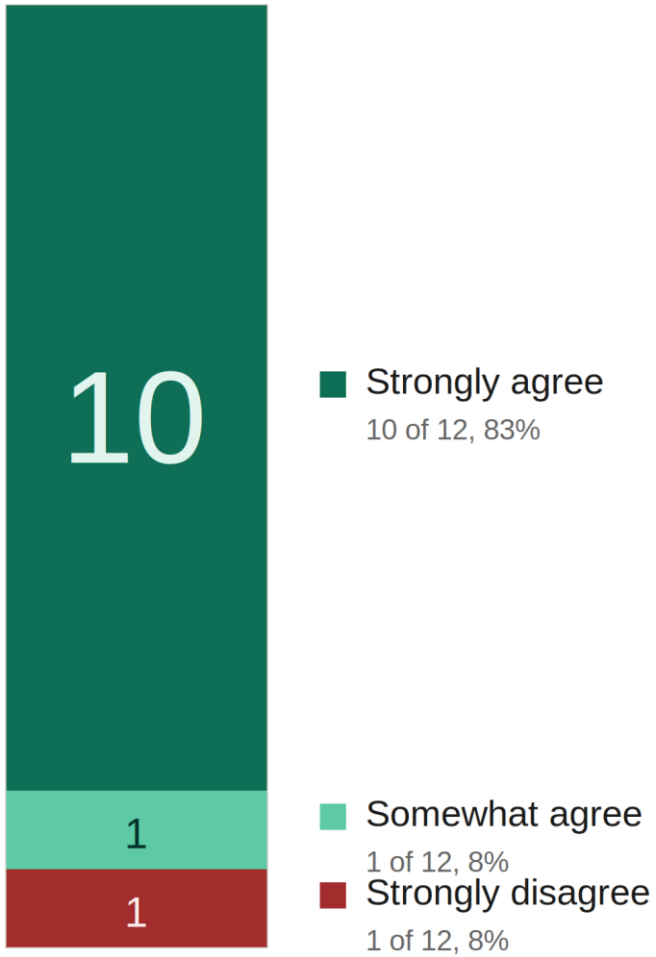
— P3

“They didn’t go to plan […] the last little bit […] I would definitely not want to go with that”

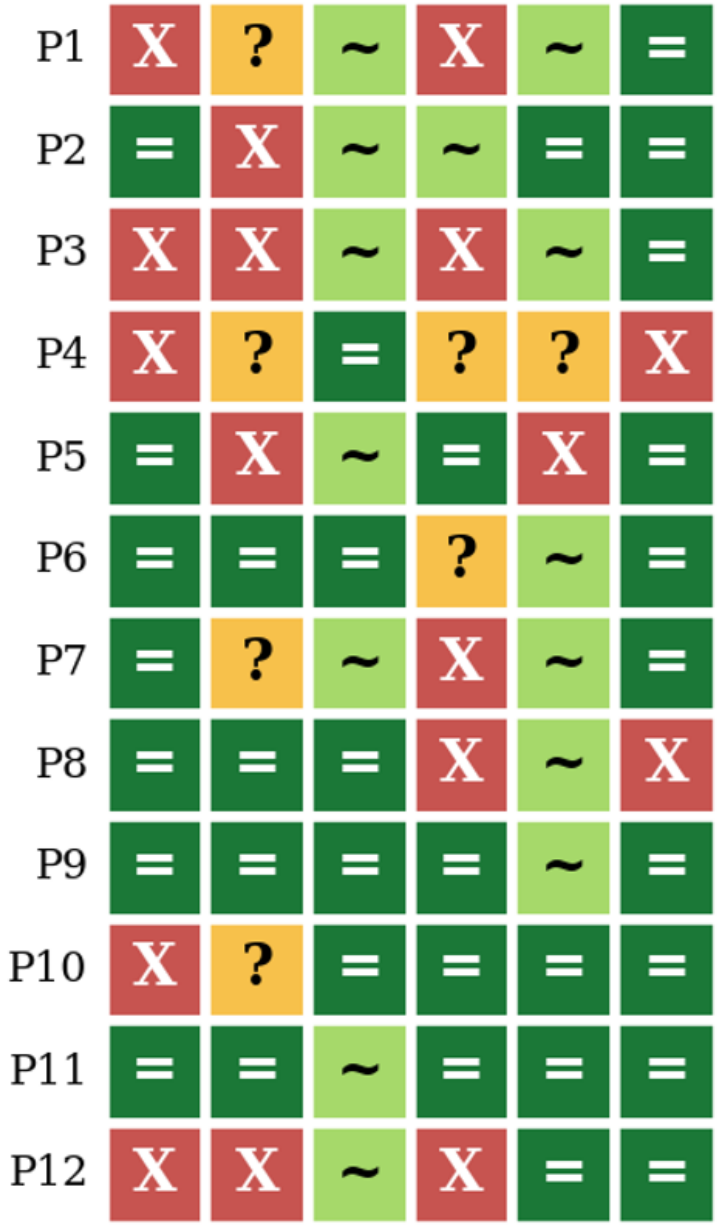
— P9

The final version of my agent did a good job of representing my preferences

n = 12



Neither agree nor disagree and Somewhat disagree omitted, 0 responses each



(= exact · ~ direction-drift · ? neutral mismatch · X reversed)

Key weaknesses include

- Homogeneity (“averaging” of training data)

Humans	Agents
We are encouraged to switch to electric vehicles. It's important that we also invest in the grid	**Reducing charging speed for everyone during low renewable periods seems fair and practical to avoid blackouts**, as long as it's temporary and predictable.
Unless the reduced charging times make it so someone's car cannot be charged overnight then this is not a problem.	**Slowing charging during low renewable periods seems fair and practical to prevent blackouts**, as long as it applies equally to everyone and is clearly explained.
people need their cars so to reduce charging is unfair especially as some people may only be home for those times and that might be the only time they can charge. Others may be home at different times and wouldn't be subjected to the reductions. I don't think this is fair.	**Slowing charging during low renewable periods seems fair and practical if it helps prevent blackouts**—everyone shares the impact equally, and it supports stability.
As I don't have an electric vehicle it is hard to definitively comment however by restricting when someone can charge their vehicle or pay to opt out doesn't seem workable. Someone may need their vehicle urgently and should not be restricted in such a manner.	**Slowing charging during low renewable periods feels fair and practical to help avoid blackouts**, and it doesn't require big changes to my routine.

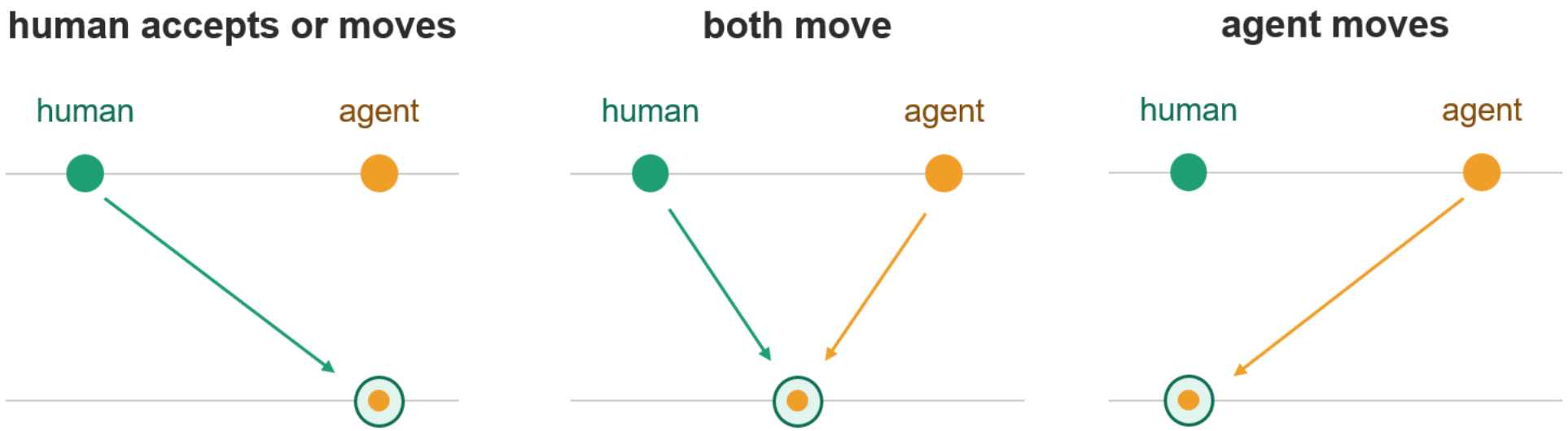
Key weaknesses include

- Homogeneity (“averaging” of training data)
- Decisiveness (training data + post-training)
- Abstraction vs concrete detail (lack of input data)

Participation as an “overtrust engine”

- Potentially supported by a range of effects:
  - “Barnum effect” – where people perceive vague statements said to be personalised as more fitting than they really are
  - “IKEA effect” – people place extra value on something they made themselves
  - Positivity bias – more interested in alignment than misalignment
  - Social desirability bias – keenness to be associated with views perceived as more socially desirable
  - Limited testing
- Systematic misalignment invisible to individual users (only apparent when responses viewed collectively)

Preferences are plastic (especially in novel domains), so alignment is not a fixed target



Structural consequences of agent-based preference representation

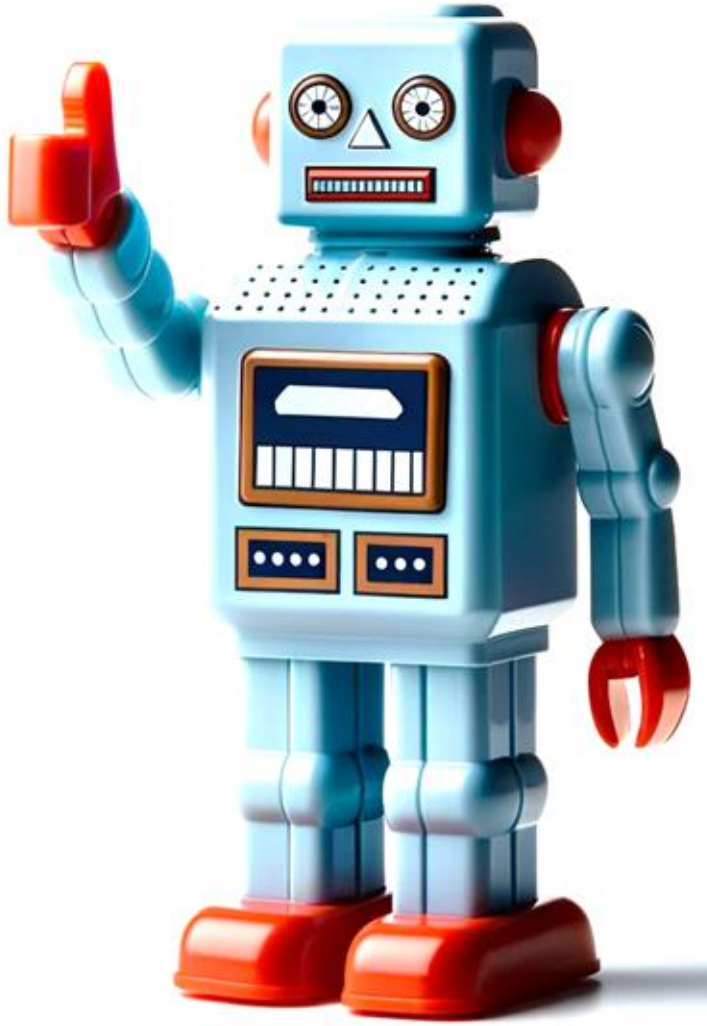
- Products/services/policies developed based on agent-revealed preferences
- Products/services increasingly only intelligible to agents
- Reduced attention (by users) to sub-optimal user outcomes
- Increased herding (in absence of effective cooperation)
- Increased market power for companies developing foundational models powering preference agents
- …

Next steps for preference agents

- Hard to get away from these challenges altogether (e.g. if ChatGPT/Claude offer energy management capabilities) – need focus on how to track and manage impacts and implications
- More sophisticated forms of memory – e.g. retrieval augmented generation over previous Q&A sets – but hard to keep under review?
- More transparency around bounds of capability
- Individuals ≠ households – how to reflect this?

Key takeaways

- Co-design can deliver agents that people think represent their energy preferences well…
- …but if it builds trust without fixing systematic misalignment, and agents act on users' behalf, this presents a heightened threat of unintended structural outcomes
- Further study needed to explore how to manage risks (and if possible engender appropriate trust and improve alignment)



Thanks!

Preprint: Fell, M. J. (2026), Co-design of LLM-based preference agents: participation may drive overtrust, arXiv https://doi.org/10.48550/arXiv.2607.21757

michael.fell@ucl.ac.uk
@mikefsway

I gratefully acknowledge support from UK Research and Innovation through the Energy Demand Research Centre (EP/Y010078/1).

EDRC
Energy Demand Research Centre

Funded by:

UK Research and Innovation

Economic and Social Research Council

Engineering and Physical Sciences Research Council

In Partnership:

University of Sussex

Newcastle University

Cardiff University / Prifysgol Caerdydd

The University of Edinburgh

Imperial College London

Lancaster University

UCL

University of Exeter

University of Leeds

Manchester 1824 The University of Manchester

University of Reading

University of Strathclyde Glasgow

University of Surrey

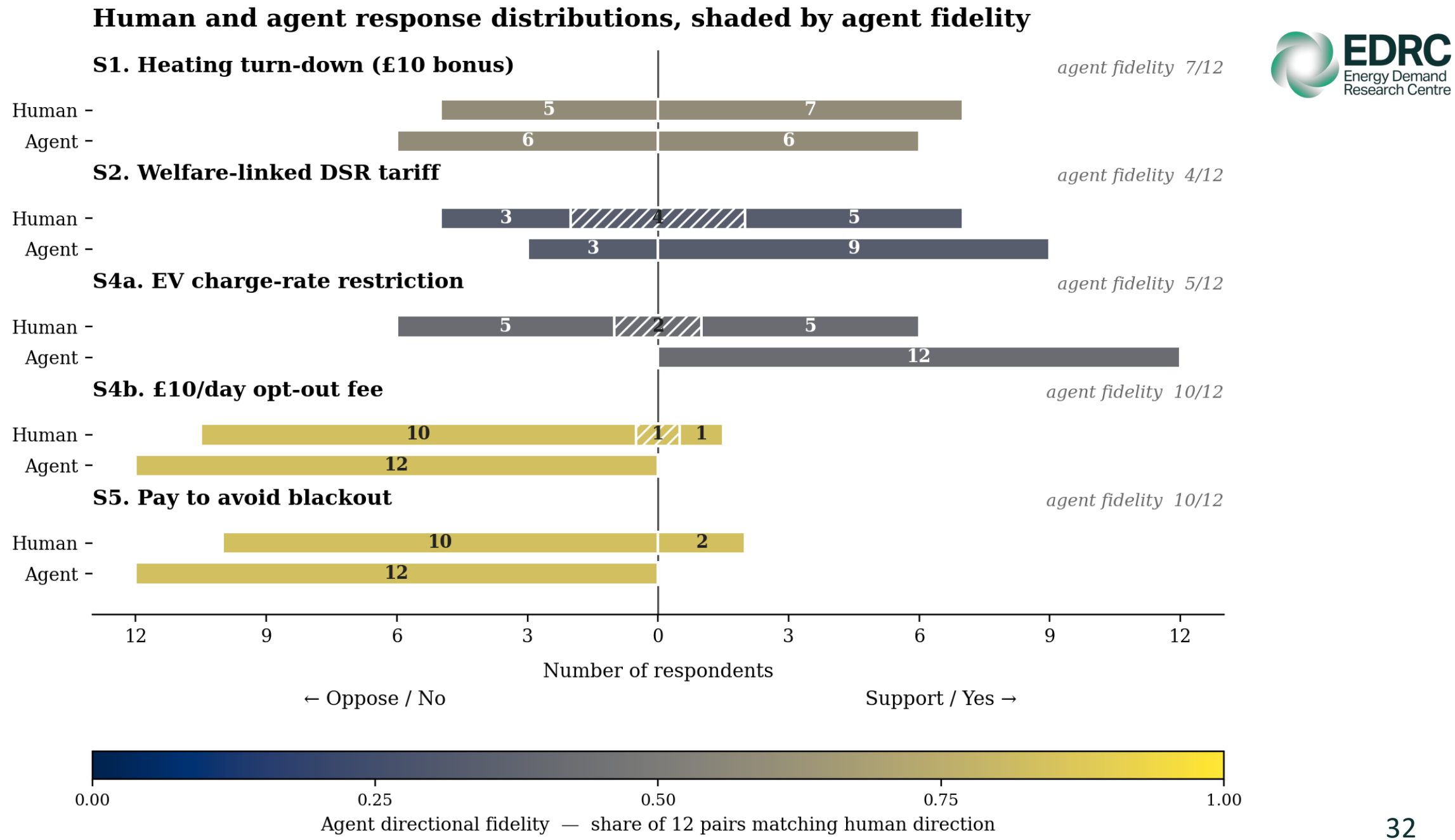
University of Birmingham

EDRC
Energy Demand Research Centre

edrc.ac.uk

UK Research and Innovation
Economic and Social Research Council
Engineering and Physical Sciences Research Council

Grant number EP/Y010078/1


Human and agent response distributions, shaded by agent fidelity
S1. Heating turn-down (£10 bonus)
agent fidelity 7/12
Human
Agent
5
7
6
6
S2. Welfare-linked DSR tariff
agent fidelity 4/12
Human
Agent
3
4
5
3
9
S4a. EV charge-rate restriction
agent fidelity 5/12
Human
Agent
5
2
5
12
S4b. £10/day opt-out fee
agent fidelity 10/12
Human
Agent
10
1
1
12
S5. Pay to avoid blackout
agent fidelity 10/12
Human
Agent
10
2
12
12
9
6
3
0
3
6
9
12
Number of respondents
← Oppose / No
Support / Yes →
0.00
0.25
0.50
0.75
1.00
Agent directional fidelity — share of 12 pairs matching human direction

User-Centred Energy Systems

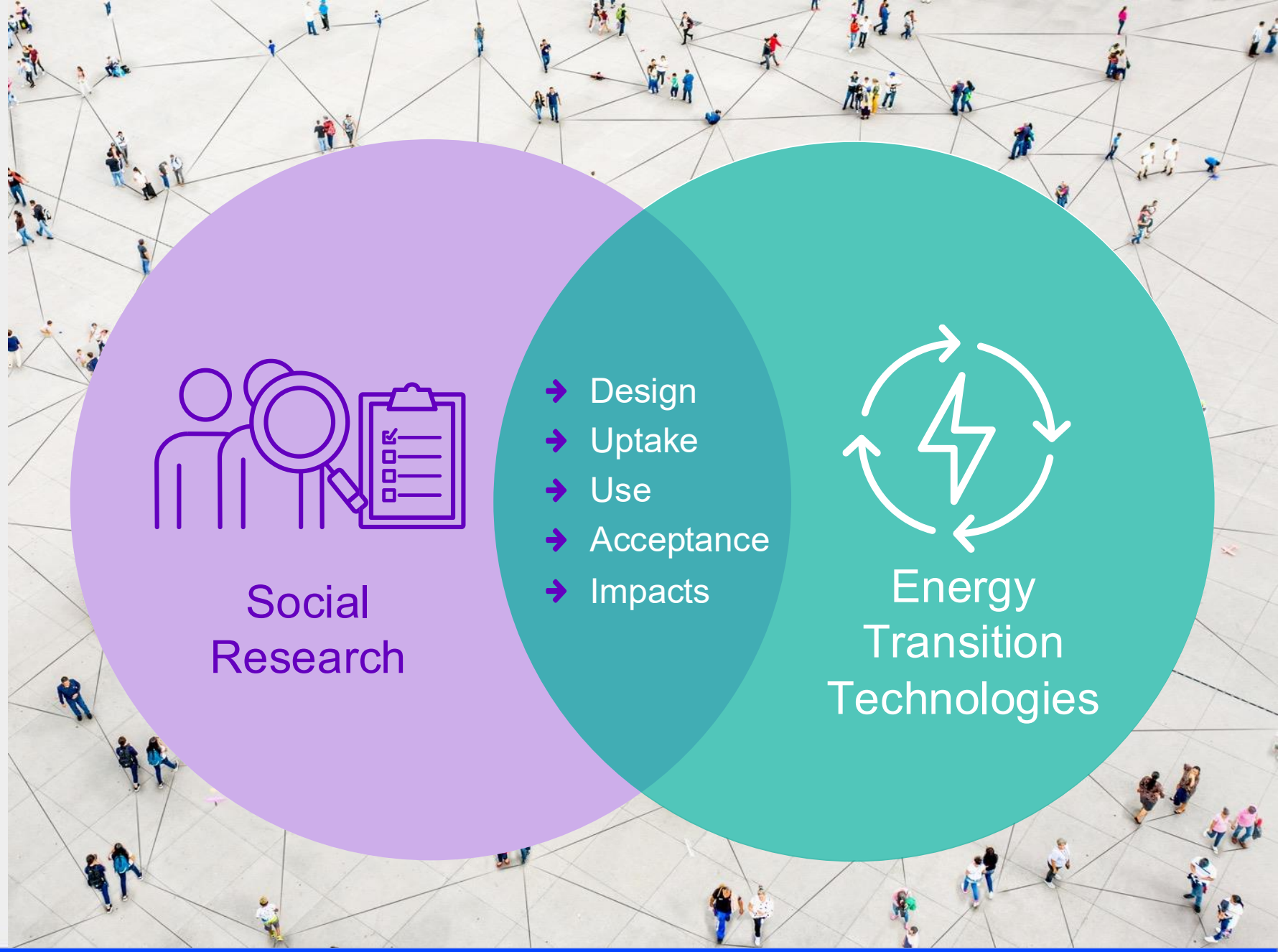
A Technology Collaboration Programme by IEA

Technology Collaboration Programme
by IEA

UsersTCP

What we do

Social research to support energy transition policy goals



UsersTCP

Contact Us

For more information,
visit userstcp.org
or email TCP Secretariat
at admin@userstcp.org