cheqd

Entangling Design

An Experience Report on Co-Creative Socio-Technical Design for Verifiable AI Agents

Entangling Design

An Experience Report on Co-Creative Socio-Technical Design for Verifiable AI

June 2025

Author: Nicky Hickman

1 Executive Summary

This report documents the design and implementation of an experimental socio-technical design methodology developed during the *Verifiable AI with Self-Sovereign Identity* initiative, led by cheqd in collaboration with SPRITE+[1], DoraHacks[2] and Verida[3]. Through a series of Social Design Jams (SDJs), the project brought together technologists, social scientists, and practitioners to explore what it means to design Artificial Intelligence (AI) agents or systems that are not only trustworthy but ethically and ecologically grounded.

Using a participatory process that combined storytelling, personas, speculative fiction, and co-creation, participants developed two complementary *Relationship Guides*, one human-centred, the other ecocentric and entangled. These outputs helped reframe AI-human relationships through diverse cultural and ethical lenses, including non-anthropocentric and posthumanist perspectives.

Key innovations included the use of fictional personas (such as Mother Earth), a reflexive and adaptive design structure, and the conceptual development of AI-powered *persona reps* as scalable proxies for participatory input. This report offers insights into the methodological challenges and successes of the process and proposes new directions for embedding relational, ethical, and ecological thinking into the design of trustworthy AI systems.

[1] https://www.spritehub.org/
[2] https://dorahacks.io/
[3] https://www.verida.network/

2 Table of Contents

3 Introduction

The emergence of agentic AI systems, tools that can act semi-autonomously and represent human or organisational intent, raises urgent questions about how we design for trust, responsibility, and long-term societal and environmental impact. While frameworks like Asimov's Laws of Robotics offered early speculative ethics, they fall short in real-world digital contexts, where fast-paced, commercially-driven development often prioritises technical functionality over social values.

This report presents a novel co-creative methodology for socio-technical design that foregrounds relational ethics, trust infrastructure, and participatory design thinking in the development of verifiable AI agents. Piloted alongside the cheqd.io *Verifiable AI* hackathon. in partnership with DoraHacks, they ran a technical hackathon focused on the intersection between self-sovereign identity, decentralised identity, verifiable credentials and AI, what cheqd have termed *Verifiable AI*[4]**.** The hackathon had two tracks, one designing for AI agents, the other covering other uses of AI for trustworthy AI systems (e.g. content credentials and verifiable data sources). cheqd's aim was to kickstart development of trust infrastructure for AI to combat the likes of deepfakes, misinformation and generally dangerous exposure to systems which do not have sufficient guardrails. The 'social design process' combined established design tools such as personas and design fiction, with transdisciplinary insights from anthropology, ecology, and systems thinking.

The objective was twofold: to surface new design principles for trustworthy AI relationships, and to evaluate whether participatory techniques could scale and adapt to the speed of contemporary digital development. Through a series of Social Design Jams (SDJs)[5], we explored the design of AI systems not just as tools, but as relational actors embedded in social and ecological networks. The process yielded two *Relationship Guides* and a set of practical, conceptual, and methodological insights with implications for design, governance, and innovation strategy.

4 Background and Rationale

[4] See https://cheqd.io/solutions/use-cases/verifiable-ai/ for further information
[5] See Methodology below, for details on what Social Design Jams are.

Since Asimov's 3 Laws of Robotics were published in 1947, academics, ethicists and policymakers have sought to define principles and rules about the responsible use of AI.[6] Yet how can we realistically implement them in the real world where commercial incentives and risk-based decision-making are paramount and the reality of the product design and development process prioritizes minimum viable product?

Starting from the premise that this calls for a new way of thinking about how to design for digital systems that are increasingly autonomous yet deeply interconnected, Interdependent and interrelated with both human society and the natural world; this socio-technical design process brought together social scientists with technologists to co-design a practical set of guidelines that can be used in the heat of product development.

The process was experimental but drew on established methods of participatory research and design. It is hoped that this experience report will provide insights for others who are using participatory or co-creative practices in their work.

5 Methodology Overview

The methodology drew on several participatory techniques used in research and commercial product development.

5.1 Methods

Design Thinking: "*a mindset and approach to problem-solving and innovation anchored around human-centered design.*"[7] Design thinking is common practice in technology development and focuses on solving customer, user or technical problems through questioning and critical thinking.

Persona: These are fictional characters normally developed based on market and customer research or other data. Personas allow product development teams to design for customers rather than for themselves and give a common basis for understanding and articulating customer and user needs.

[6] See for example,Corrêa, N., et al., *Worldwide AI Ethics: a review of 200 guidelines and recommendations for AI governance*, (19 Feb 2024), https://doi.org/10.1016/j.patter.2023.100857

[7] Han, E., "What is Design Thinking and Why is it Important?", (18 January 2022), Harvard Business Review. https://online.hbs.edu/blog/post/what-is-design-thinking [Accessed 05/06/2025]

Design Fiction: *"Design fiction is the deliberate use of diegetic[8] prototypes to suspend disbelief about change."*[9] Design fiction draws on storytelling, prototyping and science fiction to engage diverse stakeholders in product design for near futures at low cost.

Design Jam: *"a collaborative brainstorming activity or event, geared towards generating solutions in a fun and creative environment."*[10] Useful for removing barriers and constraints to ideation and engaging diverse stakeholders in participatory design practice. In our process we called these SDJ's (Social Design Jams).

5.2 Roles in the Process

During the course of the experiment we changed our approach due to a number of factors detailed below, so that our methodology was refined. There were four core roles in this iterative and reflexive process.

1) The **outcome owner**: Defines the outputs and objectives of the process, similar to a Key Customer in agile development practices.
2) The **curator**: Guides the process and participants towards its goals, prepares events, analyses feedback. This role is similar to the Product Owner in agile development practices.
3) **Facilitators:** Facilitates events (i.e. workshops and Social Design Jams).
4) **Participants:** Take part in events providing synchronous and asynchronous feedback.

5.3 Pattern of Activities

The overall pattern of activities in the core iterative and reflexive process is very similar to agile development lifecycles.

---

[8] ***Diegetic:*** existing or occurring within the world of a narrative rather than as something external to that world. (Merriam Webster Dictionary https://www.merriam-webster.com/dictionary/diegetic)

[9] Sterling, B., *Patently untrue: fleshy defibrillators and synchronised baseball are changing the future,* (11 October 2013), Wired Magazine https://www.wired.com/story/patently-untrue/ [Accessed 05/06/2025]

[10] Participedia, https://participedia.net/method/4620 [accessed 05/06/2025]

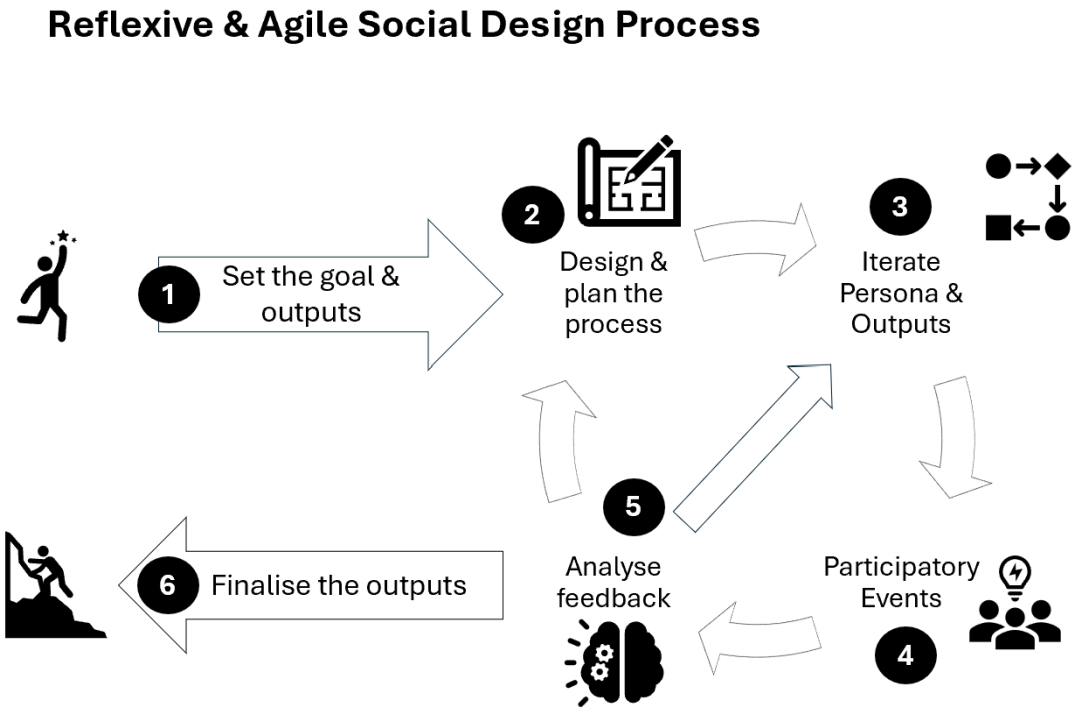


Figure 1 - Overview of basic iterative process

1) The outcome owner **set the goal and expected outputs**

2) The curator **designed the process, prepared draft** personas and invited participants from the industry and academia with expertise in social sciences and digital trust to an initial Scene-Setting Workshop.

3) **The Scene-Setting Workshop** was structured as follows:

    a) **Subject matter grounding:** Participants were introduced to AI, the key technical and market trends and the problem space of trustworthy AI.

    b) **Science Fiction stories:** Participants were asked to read or watch their own choice of two science fiction stories about autonomous machines ideally, they were asked to watch two with different perspectives on the roles of autonomous machines (i.e. dystopian and utopian)[11]. The first part of the workshop was then sharing their thinking on what differentiated these two types of imagined AI agents, especially the nature and intent of the machine's

[11] See Annexe 1 for example stories that were suggested.

creators. During this part of the workshop we generated questions that needed to be answered for persona later in the SDJs.

c) Voting on the priority **questions to answer** in the SDJ and refining the persona

4) **Strawman Outputs:** Based on analysis of the Scene-Setting Workshop the curator then creates strawman outputs, in our case this was two alternative *Relationship Guides for AI Agents and their Creators.*

5) **Social Design Jams (SDJ's):** We carried out four of these in our experiment, however depending on the goals and expected outputs of the process more or less of these may be required. In these a facilitator introduces the topic (sets the scene) in plenary, then facilitators work with a maximum (ideally) of five participants per breakout group. They select a persona and a strawman output and then brainstorm around the output from the perspective of their persona. The facilitator helps them document their work. They return to plenary and play back the outputs from their breakout group.

   In our experiment, two of the SDJ's were virtual and two were face-to-face. One of each was broadly with either industry professionals who were technically biased, or with participants from academia mainly the social sciences. There was roughly a 60:40 split between people identifying as male vs female, and the participants were mainly from the global north with ~30% of participants with black or ethnic minority heritage

6) The curator then analysed the feedback from the workshop and the SDJs and **produced the final outputs.**

6 Case Context: Verifiable AI with SSI

The pilot took place alongside a "Verifiable AI" hackathon run by cheqd.io, in partnership with SPRITE+, Dorahacks and Verida. The hackathon invited developers to build AI solutions that leveraged decentralised or self-sovereign identity technologies for trust-building functions such as content credentials for verifying provenance and authenticity of content, digital identity for proof of personhood and Know Your AI/Agent to know who you are dealing with online, and other applications all supporting trustworthy AI, with traceable and verifiable supply chains including factors such as data provenance.

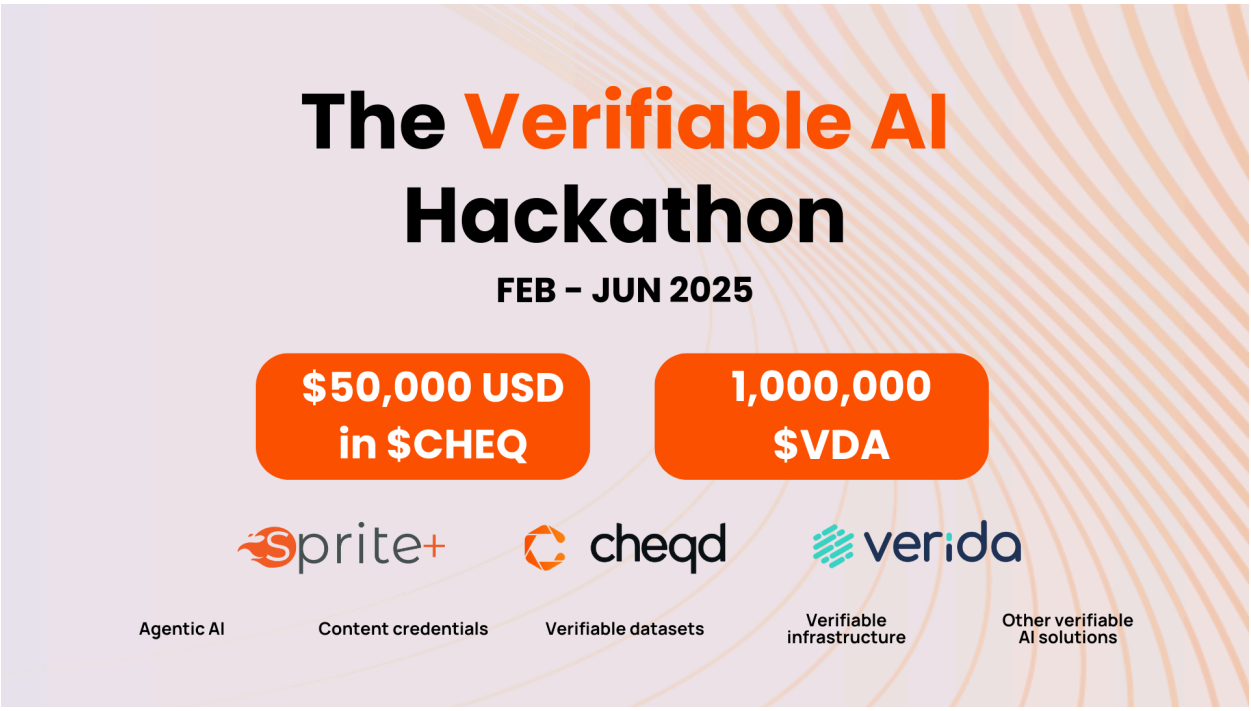


Figure 2 - Poster for the Verifiable AI Hackathon

The purpose of the Social Design track was to understand how these technical building blocks of trust could be used to strengthen a trustworthy relationship with AI systems and in particular, AI agents. We focused on AI agents because these systems are most likely to attempt to represent human counterparts or co-workers, and hence most likely to be anthropomorphised by their users.

If *"trust is a confident relationship with the unknown"*[12], then the question this process sought to answer was; "***How do you design , build and operate AI agents for healthy and trustworthy relationships with humans?"***

7 Implementation and Experience

The implementation phase of the project centred on a series of Social Design Jams (SDJs) that brought together participants from diverse disciplines to engage with ethical, relational, and ecological dimensions of AI agent design. These sessions served as both

[12] Botsman, R., "Who can you Trust?", (November 2017), Hachette, ISBN-13 9781541773677.

experimentation spaces and reflective inquiry, testing the viability of participatory methods including personas, storytelling, and speculative prompts within a fast-moving technological context. This section outlines how the methodology unfolded in practice, what adaptations were made, and what was learned from the experience of co-creating trust-oriented AI design guides.

7.1 Plan A - One Social Design Jam

Initially the process was designed to include one Social Design Jam with social scientists and service designers, (See Annexe 1). This was intended to be followed by two interactive sessions with the hackers.

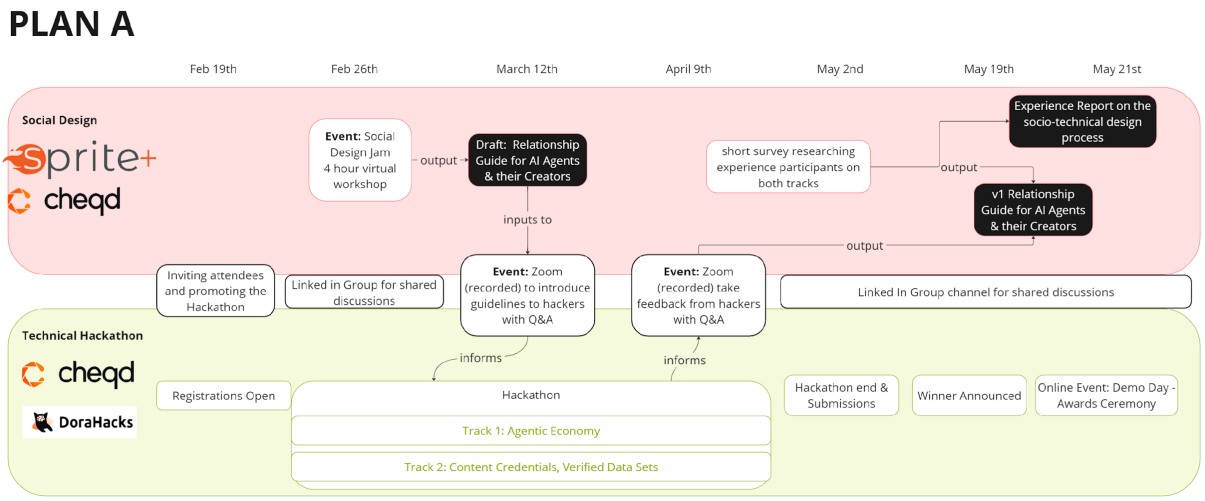


Figure 3- Initial plan for the socio-technical design process showing only one SDJ

As this was a virtual workshop, and many of the participants had never met each other before, the pre-work discussion (using science fiction stories to highlight key questions to ask for our persona), and introductions became extremely important for level setting and establishing good working relationships.

We included an overview of the technical space in terms of state of the art on AI and the issues of trust. Of particular interest were questions of authorisation (permissions), authenticity (e.g. of data and its provenance), and authority (flows of responsibility, accountability and liability) for decisions and actions that AI agents may take. This led to a discussion that challenged our use of the term AI agents, participants highlighted that the term agent

suggested agency and this was a contested term. We were encouraged to use the term 'autonomous systems' instead.

Participants were then introduced to the persona and the ensuing discussion highlighted important other factors in terms of relationships that we had not previously considered due to our focus on the application of particular technologies to the human-AI relationships. These were rules, power, anthropomorphism of AI systems and domains such as gaming where AI systems were considered 'custodians of truth'.

Importantly participants in the workshop highlighted the importance of environmental sustainability and impacts of AI in our thinking about how to build ethical AI systems.

The second half of the workshop where we applied the questions to the persona effectively became a trial run for subsequent Social Design Jams. Following the workshop we requested feedback from participants which included recommendations that we follow up with a second Social Design Jam in order to fully explore relationships for the persona.

7.2 Plan B - Set the scene and multiple Social Design Jams

Plan B came about partly as a result of recommendations from participants in the first workshop, and partly because the engagement from hackers was very low and it was decided that interactive sessions with them would not have produced the desired results. We therefore sought other technical audiences and decided we would run SDJ's with them as an alternative.

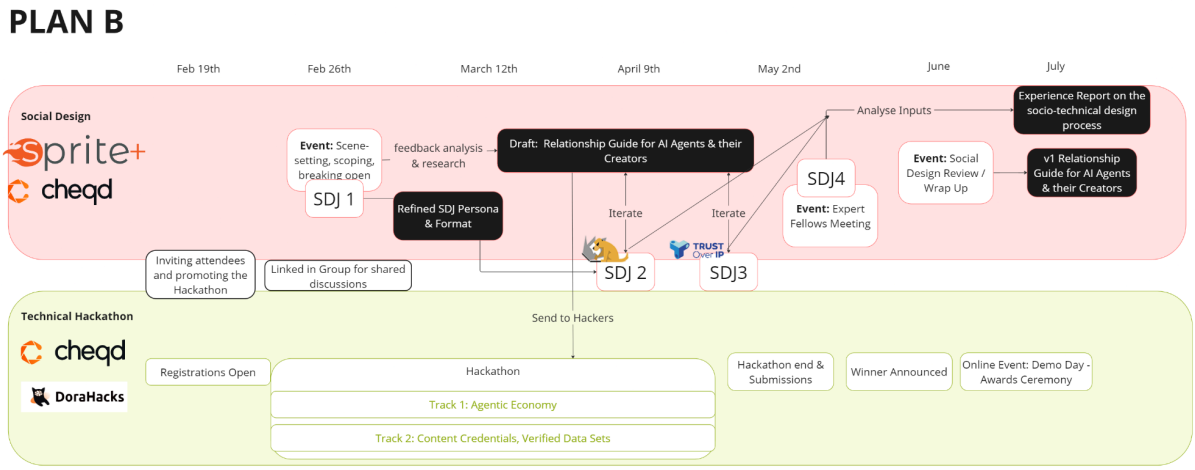


Figure 4 - Revised plan containing a Scene Setting Workshop and additional design jams

The revised plan and process was more iterative and used the Social Design Jam (SDJ) format instead of online workshops with technical audiences at the Internet Identity Workshop[13] (SDJ 2) and within the AI and Human Trust Working Group at Trust over IP Foundation[14] (SDJ 3). As the hackathon was extended to June, and SPRITE+ already had an Expert Fellows Meeting scheduled on the subject of human - AI relationships, we took the opportunity to also run a short SDJ 4 at this face-to-face meeting. SDJ's 2-4 were 1 hour long.

We used the output from SDJ 1 (part of the scene-setting workshop) to draft the outputs, refine the persona and to develop an SDJ Pack (See Annexe 2, Participant Guide from the SDJ Pack). This enabled different facilitators to run their own SDJs. The feedback from this first session also led to some important changes in the persona which were enriched with insights from participants, and led to the creation of not one, but two alternative *Relationship Guides*. We created a human-centred guide, and an entangled guide that offered a non-anthropocentric view of human-AI relationships. We also added a fourth persona which was *mother earth.*

Participants in all the SDJ's were asked to add their notes and comments to shared online documents of the relationship guides and of the persona. Feedback from the SDJ's was used to refine and improve the *Relationship Guides* as final output from the process.

These practical engagements not only surfaced design tensions and methodological challenges but also revealed deeper patterns of value, concern, and aspiration, insights that informed the development of the *Relationship Guides* and reflections on future experimentation and inquiry.

8 Key Learnings on the Methodology

Overall participants responded positively to the interdisciplinary format and ethical framing. In particular all participants appreciated the opportunity to explore the issues from the perspective of persona and storytelling rather than relying on individuals' own experiences. Unexpectedly, it was more technical groups who most appreciated the social design methodology and use of persona and storytelling.

As with other design jam approaches, participants had fun and enjoyed the experience. Being able to influence the social design process as well as the outputs also seemed to lead to greater engagement, the idea that everyone was experimenting together removed pressure on specific outputs from each SDJ.

[13] https://internetidentityworkshop.com/
[14] https://www.trustoverip.org/

Key challenges were:

- Lack of time to work through a full *Relationship Guide*, most of the SDJ's were only 1 hour long, and this was not really time to give everyone space to contribute or to get through all the facets in each *Guide.*
- Virtual SDJ's did not work as well as face-to-face events, even where the participants knew each other prior to the event, as in the case of the Trust over IP group.
- Although the SDJ Pack that was developed by the curator was self-explanatory, it was too long and dense so in future simplified cards for persona and SDJ's more similar to the presentation format (see Annexe 2) would work better.
- Engagement from the hackers in the hackathon and alignment with their schedule did not work well, in future it would be better to start the social design track well ahead of the development effort so that there could be direct input or requirements specifications into the development teams.
- We did not gain informed consent from participants in the SDJ's, this was a major oversight and meant that we could not use the output from one of the SDJ's. In a more formalised and structured process this must be included.
- Engaging the diverse groups into a shared discussion. Originally we had intended to use Linkedin Groups to build a conversation and community around the concept of human-AI society. Linkedin was selected as an accessible platform for both academia and industry including technologists. However, this neutral territory was a no-man's land between academia who used email most comfortably, and industry who were used to platforms like Slack and Discord.

9 Novelty and Contribution to Methodology

This experience report makes several novel contributions to the field of participatory and socio-technical design particularly in its application to the emerging domain of Verifiable AI and autonomous agents. While grounded in established co-design practices such as personas, design fiction, and design jams, the process innovatively extended these methods to address new questions of trust, agency, and ecological interdependence in human-AI systems.

Expanding the Scope of Participation

Traditional participatory design focuses on human users as central stakeholders. This methodology pushed the boundaries by incorporating non-human entities—such as

ecosystems and *machines*—as active participants in the design conversation. The creation of the Mother Earth persona introduced an *eco-centric and post humanist perspective*, moving beyond anthropocentric norms toward *entangled design*. This aligns with recent theoretical trends in critical post humanist participatory design, but few practical design methodologies have operationalised these ideas as successfully.

Dual Relationship Guides: Human-Centred and Entangled

The development of two distinct yet complementary *Relationship Guides;* one rooted in human-centred design and the other in entangled, post humanist systems thinking is methodologically significant. It demonstrates how different epistemological frames (individualist vs. systemic, Western vs. Indigenous-informed) can coexist and inform relational technology design. This dual-guide approach challenges assumptions of design universality and provides pluralistic tools that teams can adapt to diverse cultural or environmental contexts. (e.g., Escobar, A., 2018)

Methodological Reflexivity and Adaptive Framing

The shift from a single workshop to a distributed, iterative series of Social Design Jams (SDJs) exemplifies a high degree of reflexivity and responsiveness which are core values in participatory action research but often lacking in structured co-design programs. The ability to evolve the process in response to stakeholder feedback (e.g., extending beyond the hackathon, adjusting facilitation materials, refining personas) indicates a living methodology, one that adapts to the real-world complexities of socio-technical system development.

Ethical Speculation as Practical Design Input

The integration of science fiction prototyping and narrative prompts to stimulate ethical and practical reflection on AI–human relationships is not only novel but necessary in a context where regulatory foresight lags behind technological advancement. Rather than remaining in the realm of abstract speculation, this project successfully channelled these fictional prompts into actionable questions and relational criteria used in SDJ's, bridging speculative inquiry with applied design.

This contribution resonates with the literature's call for *expanded, reflexive, and systemic approaches* to participatory design in the age of autonomous systems. (e.g.Delgado, F., 2023; McCarthy, P., 2015). It provides both a theoretical provocation and a practical

framework for embedding values, diverse voices, and ecological reasoning into the core of AI design practices.

10 Reflections and Recommendations

The methodology described above started with three root questions:

- How to design for socio-technical systems?
- How to keep the customer in the room in a simple, low-cost way?
- How to design AI systems that are verifiably trustworthy?

The results of the exploration sparked two branch questions for future experimentation and enquiry:

- How to make participatory design scalable and cost-effective by using AI agents to represent persona, machines and nature in the continuous design, development and operation of cyber-physical systems?
- How could design practice effectively nudge our AI trajectory away from the interests of now and towards planetary well-being for future generations by rethinking our relationships with machines and the natural world?

10.1 Socio-Technical-System Roots

Design for socio-technical systems

Although this is often cited as an approach, it is rare to see it really in practice. We have privacy by design, ethical design, respectful design, inclusive design, security, safety and human-centered design but what does this mean on the shop floor with software and digital development teams? All too often, despite the widespread adoption of agile in software development practice, the social side of socio-technical design is still at the back of the process; left to user-experience designers, compliance officers and ultimately, customer services advisors. *What would happen if we front-loaded the social design and drew on the expertise and insights of anthropologists, psychologists, sociologists and other social scientists?*

Keep the customer in the room

Participatory and co-creative design practices in many domains lead to better outcomes and products. Developing any product or service should start and end with its users or customers, focusing on outcomes that are both profitable and purposeful, promoting commercial, social and environmental wellbeing. However much of the current developments in both digital trust and AI remain the preserve of a technical elite. Participatory design is also costly, time-consuming and complex to manage with many pitfalls, especially to maintain on an ongoing basis. *How far could persona take us in a proxy of participation for diverse stakeholders?*

In this case, the methodology worked to a degree, however in future it should be refined and improved:

- Account for more time, more real-world SDJ's
- Speed dating - Scaling relational design for socio-technical systems with *persona reps* and *machine reps.*

The breathtaking speed of change and technological advancement that the current AI era has brought upon us is both exhilarating and terrifying. Whilst there is hype aplenty, particularly around reasoning and artificial general intelligence, there is also not going to be a trough of disillusionment in the same way as there may be for other technologies as AI is a general purpose technology and it's here to stay. Given the many human and environmental harms that arise from misuse of powerful technologies, there is new urgency in finding ways of incorporating humans not somewhere in the loop (➰), but everywhere in the lemniscate[15] (∞).

This social design process showed that participatory design and the use of techniques such as fictional speculation could indeed accelerate and unlock our thinking on how humans would interact with AI systems in the near future. It also offered an insight into how we could do this in practice at speed and scale in rapid, iterative and agile research, product, policy or standards design and development processes.

AI agents as *reps*

This exploration and the curation of this process involved extensive use of ChatGPT, both as research assistant, critic and co-author of the *Relationship Guides*, without it, the work would have taken many months and more resources, but AI was used as an occasional tool rather than as a constant collaborator. A next step in terms of experimenting with this

[15] The symbol for infinity

approach is to create AI agents to represent persona, informed by the data about customers, users or domain experts such as regulators, social scientists or ethical hackers. These *persona reps* could cooperate with, and inform designers, product managers and engineers throughout the development process.

This would have several advantages over other methods of participatory design including cost reduction and increased speed, scale, as well as naivety and diversity of participants. Imagine being able to run UAT's past a bunch of *persona reps* that test new features as part of a devops process? Or being able to get *persona reps* to prioritize your backlog by voting. Imagine if those *persona reps'* behaviours and perspectives were dynamically informed by the latest market research or behavioural data from the users, customers or the experts they represent?

On the flipside, agents could also represent machines or AI systems, UX or service designers could interrogate these *machine reps* in real time in terms of the cost and feasibility of desired features or customer experience flows. Cooperating together, different *machine reps* from distributed and diverse systems could inform strategic decision-making and reduce the time, risk and cost of complex systems development by providing predictive insights on the capabilities and constraints of the technical part of the socio-technical system.

10.2 From Socio-Technical to Eco-Socio-Technical Systems

At the outset, our social design process was very clearly rooted in design thinking and a western, individualist, anthropocentric worldview. Our persona were all consumers and our frame of reference was largely personal AI systems working on behalf of customers and companies interacting with each other. However many AI systems include IoT (Internet of Things) devices and have application domains such as agriculture, meteorology, or manufacturing which do not include many human-AI interactions at all. Then, during the landscaping workshop there was a detailed discussion about the environmental impact of AI systems, also a concern for policy-makers and for future generations.

These factors led to the creation of a second *Relationship Guide*, that took an entangled, eco-centric or post-humanist worldview as its starting point. Whereas the first *Guide* drew on psychology, sociology and anthropology for its relationship facets, the second, entangled *Guide* drew on indigenous wisdom, systems thinking and ecology. In this entangled

worldview, dominant in the global south (Escobar, 2018), individual humans are just one entity in a co-dependent web of relationships that includes social groups, machines and the natural world.

This brings us to a third type of entity that agents can represent, *nature reps* can be present in decision-making and design. Imagine designing a new building management system for a new development close to a river, perhaps even a sacred river with personhood, like the Whanganui River in New Zealand. As you set the location of sensors and calibrate the thermostats or position the wind turbine, imagine if you could consult *nature reps* that knew about water flow rates, seasonal changes or could represent the migratory path of birds in the area? Use of digital twins for environmental monitoring and as representatives of entities in the natural world are already in use. Could we use AI to also give them a voice, and thus enable them to meaningfully integrate into the continuous flow of socio-technical systems design, development and operation? This could be a step change towards not just sustainability, but also restoration and regeneration. Perhaps most importantly, it begins to rebalance the dominance of a global north worldview that is at best only half the real market, and at worst is a new form of cultural imperialism. Further exploration and experimentation is required to understand the creation and use of *reps*, together with their ethical implications. Whilst they are not proposed as a substitute for real human participation, they could be a support.

Persona, nature and *machine reps* that embody the needs, rights, and perspectives of various stakeholders (including future generations and nature) could offer a speculative but pragmatic response to the challenge of scaling participatory design. In practice, these could help embed ongoing stakeholder input into rapid design cycles, policy co-creation, and agile product development processes, representing a step-change in how participatory design might evolve with the support of AI itself.

10.3 Jam today – Practical Recommendations for Practitioners

The overall goal of our process was to create *Relationship Guides* that were of practical use to designers and developers working with AI now. None of the methods used are unusual, costly or difficult to implement, furthermore the process is highly adaptive to prevailing circumstances and resources. Example uses of this approach that you can implement today are:

1. **Use Personas Early and Often**
   Develop and validate personas (including non-human stakeholders) at project kick-off to surface relational, ethical, and ecological considerations before technical decisions are locked in.

2. **Host Micro Social Design Jams**
   Integrate short, focused Social Design Jams into agile sprints to gather multidisciplinary input on features, risks, and trust factors in real time.

3. **Use or adapt our *"Relationship Guides"***
   Use our lightweight templates to define expected behaviours, accountability flows, and trust indicators for AI systems.

4. **Appoint a Curator**
   Assign a team member to serve as curator, not just a project manager, to maintain ethical focus, prepare materials, and synthesise design reflections throughout the product lifecycle.

5. **Prototype with Fiction and Futures**
   Incorporate storytelling or speculative prompts (e.g., "what would your AI agent do in a crisis?") into design reviews to uncover overlooked risks and assumptions.

11 Conclusion

This experience report has shown that participatory, co-creative, and narrative-based design methods can meaningfully inform the development of AI systems that are socially and ecologically trustworthy. Through interdisciplinary collaboration and iterative experimentation, we demonstrated how social values such as care, agency, consent, and ecological stewardship can be operationalized in the design of autonomous systems.

The creation of dual *Relationship Guides*, one grounded in human-centered design, the other in entangled, post humanist thinking, marks a methodological contribution to socio-technical design. Moreover, the concept of AI-powered persona and machine representatives' points to a new frontier for participatory scalability, where AI could act not only as a design subject but also as a tool for ongoing engagement and feedback.

Yet the process also revealed limitations: challenges in sustaining participation, misalignment with hackathon timelines, and the ethical risks of using personas as proxies for

lived experience. Future work should explore how to embed these practices more deeply into agile development, policy design, and research.

In a world increasingly mediated by AI, this report argues for a shift in design paradigms, one that treats humans, machines, and nature not as separate actors but as co-constitutive agents in a shared, entangled future.

12 Acknowledgments

This project has been co-funded by SPRITE+. SPRITE+ is the UK NetworkPlus for Security, Privacy, Identity, and Trust. SPRITE+ is a platform for building collaborations across the spectrum of issues relating to digital security, privacy, identity, and trust. SPRITE+ is funded by the Engineering and Physical Science Research Council (grant reference EP/W020408/1). Find out more: https://spritehub.org

We would like to thank the following individuals and organisations for their participation in the Social Design Jams.

Andrew Slack, Head of Product, Randamu (see: https://www.randa.mu/)

Ankur Banerjee, CTO, cheqd.io

Dr Bruce White, Founder and CEO, The Organization for Identity and Cultural Development (OICD) (see: https://oicd.net/)

Chikara Shimasaki, Senior Officer, Innovation, Programmes & Partnerships, The Organization for Identity and Cultural Development (OICD)

Dr Ioanna Noula, COR (Children's Online Redress) Sandbox and Visiting Fellow at University College Dublin and London School of Economics (see: https://www.corsandbox.org/)

Mark Elliot, Professor of Data Science, School of Social Sciences, University of Manchester (see: https://research.manchester.ac.uk/en/persons/mark.elliot)

Robin Pharoah, Director, Future Agenda (see: https://www.futureagenda.org/)

Sownak Roy, Senior Engineering Manager, cheqd.io

Susan Morrow, Head of R&D, Avoco Secure (see: https://www.avocoidentity.com/)

The AI and huMan trust (AIM) Working Group at Trust over IP Foundation (see: https://www.trustoverip.org/)

The Internet Identity Workshop (see: https://internetidentityworkshop.com/)

13 References

ChatGPT 4.0, used to support research, refine and summarise text.

Delgado, F., Yang, S., Madaio, M., and Yang, Q., *The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice,* (October 2023). https://doi.org/10.48550/arXiv.2310.00907

Escobar, A., *Designs for the Pluriverse: Radical Interdependence, Autonomy and the Making of Worlds,* (2018) Duke University Press, ISBN: 978-0-8223-7105-2

McCarthy, J., Wright, P., *Taking [A]part: The Politics and Aesthetics of Participation in Experience-Centered Design*, (January 2015). DOI: 10.7551/mitpress/8675.001.0001, ISBN: 9780262328098

14 Annexe 1 - Social Design Jam 27 February 2025

14.1 Purpose

The social design jam will have output a set of Agentic AI design principles and guidelines, the relationships between AI agents and their creators. Of particular interest are questions of authorisation (permissions), authenticity (e.g. of data and its provenance), and authority (flows of responsibility, accountability and liability) for decisions and actions that AI agents may take. These will be used to initiate the co-creative process with participants in the cheqd / Dora Hacks Verifiable AI hackathon.

In the jam we will answer questions such as 'what are the relationships between AI agents and the people and organisations they work for?', 'what agent controls MUST or SHOULD be built into any AI Agent by default to ensure human agency, human in the loop and risk management

to protect from human harms', 'what type of bad actors could co-opt or influence AI agents and what are the guardrails that are needed by designers to prevent manipulation?', 'how does the lifecycle of a relationship between an AI Agent and it's employer change through life's course as capacity waxes and wanes?','what happens to AI agents when you die'? 'What will make AI agents most trustworthy'?

14.2 Method & Rationale

The market and technology are moving at pace, whereas in the past we have had years, sometimes decades to consider the social and human impacts of new technologies, today we have no such privilege, therefore we borrow from industry and use an approach called 'design jam'[16] that prioritises participation, diversity and is a brainstorm.

What?	This will be a 4 hour virtual session
When?	Weds 26th February 2025 13.00 - 17.00 UCT / GMT

14.3 Pre-work

As a source of inspiration before the workshop participants are encouraged to read or watch at least one of each dystopian and utopian stories about AI Agents, here are some suggestions from Claude.AI

For Technical Focus

- Positive: Interstellar (TARS/CASE)
- Negative: 2001: A Space Odyssey (HAL 9000)
- Theme: Boundaries between autonomy and control

For Ethical Focus

- Positive: Big Hero 6 (Baymax)
- Negative: Ex Machina (Ava)
- Theme: Purpose, consciousness, and ethical boundaries

For Practical Implementation

- Positive: Iron Man (JARVIS)

[16] See: https://participedia.net/method/4620

- Negative: I, Robot (VIKI)
- Theme: Balancing capability with control

Discussion Points for Each Viewing

1. How are boundaries established and maintained?
2. What role does transparency play in the relationship?
3. How is trust built or broken?
4. What safeguards exist and are they effective?
5. How does the AI's purpose influence its behavior?
6. What role does human oversight play?
7. How does the relationship evolve over time?

14.4 Workshop Content

We will prepare three persona for groups to design for:

1 - Jim - a 14 year old gamer

2 - Sally - a retail investor and crypto trader aged 30

3 - Jethro - a 55 year old plumber and small business owner, Jethro's wife is living with early onset dementia

Each group will have <5 participants ideally from different disciplines together with a facilitator.

14.5 Social Design Jam Agenda:

Time	**Item**	**Who**
20mins	Introductions and objectives	cheqd as hosts
30mins	**Ice-breaker & identify the questions to answer:** Share stories from input fiction and key questions they raised that should be answered in the session. Place on virtual board and cluster into groups.	All
10mins	Voting on the top 7 questions each group will answer for their AI user persona.	All
15mins	break	

30 mins	In break out groups, What types of AI agents does your persona use and why, whose other agents does it need to interact with to get the job done, how do your persona's AI agents interact with each other. What are the key success criteria for your persona with respect to their agent (s).	All
15 mins	break	
45 mins	In break-out groups answer the questions for each persona and recommend design principles, UX steps or governance rules that would mitigate any risks and promote the benefits of using AI agents.	All
15 mins	break	
15 mins	Groups playback to plenary (5 mins each)	Group facilitators
30 mins	Discussion, prioritisation, gaps	All
15 mins	Next Steps & Close	cheqd

15 Annexe 2 - Participant Guide from the Social Design Pack

Participant Guide

Social Design Jam

1. **Pick a relationship guide**
   a. Human-centred
   b. Entangled
2. **Pick a persona**
   a. **Jose** - The 14-year old student and gamer (USA)
   b. **Sally** - a freelance creative and occasional crypto-trader (South Africa)
   c. **Jethro** - The Plumber & Small Business Owner (UK) and his wife Alison who is living with early-onset dementia
   d. **Mother Earth -** The Living Planet
3. Using the themes and draft principles in the relationship guide, discuss **how the AI agent should develop, build, maintain and end healthy and trustworthy relationships with its human users.**
4. **Document – changes** to the relationship guide for that particular persona, **questions** the persona would ask its AI agents, **examples** of AI agent design that realises the design principles.

Pick a Guide?

Thematic Map: Anthropocentric vs. Ecocentric AI-Human-Nature Relationships

Onboarding, offboarding, data deletion	**Lifecycle & Transitions**	Rites of passage, release, mourning, rebirth
Appeals, user rights, redress mechanisms	**Conflict & Redress**	Restorative processes, healing cycles
Personalisation, skill-building, autonomy	**Learning & Growth**	Seasonality, co-evolution, non-linear becoming
Defined roles, permissions, user overrides	**Boundaries & Autonomy**	Non-intrusion, kinship boundaries, relational ethics
User-centric safety, risk management, compliance	**Care & Responsibility**	Stewardship, distributed care, ecological wellbeing
Predictability, reliability, emotional safety	**Trust & Bonding**	Reciprocity, ritual, rhythm of relationship
Clarity, transparency, explainability	**Communication**	Attunement, symbolic cues, listening to silence

Anthropocentric Lens Ecocentric / Posthumanist Lens

Relationship Guide 1 - Human Centred

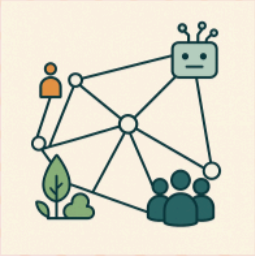
1. **Communication & Understanding** AI systems must communicate clearly, contextually, and appropriately for the human they are supporting.
2. **Trust & Emotional Bonding** AI agents should build trust over time by behaving reliably, ethically, and predictably.
3. **Care, Safety & Responsibility** AI agents must be designed with human safety and well-being as top priorities. The creator, operator, and user all carry different responsibilities. Clear accountability chains should exist.
4. **Boundaries, Control & Autonomy** AI systems must operate within clear boundaries, determined by the user and/or regulatory frameworks.
5. **Learning, Growth & Development** AI should evolve with the user—but in a controlled, observable, and ethical way.
6. **Conflict Resolution & Behaviour Management** Systems must be able to handle errors, misunderstandings, or harm in a way that restores trust and accountability. There must be mechanisms for redress, feedback, and correction.
7. **Lifecycle & Transition Planning** Relationships change. AI agents must be designed with a lifecycle in mind: from onboarding, to growth, to possible decommissioning or transfer.

https://docs.google.com/document/d/12dd4dpTEodem630icgV-Qal9NOKIfSTvUxGmsVzzxrc/edit?tab=t.0

cheqd
sprite+

Relationship Guide 2 – Entangled

1. **Communication as Attunement** AI systems should be designed to express purpose, uncertainty, and change in ways that are intelligible, respectful, and responsive to human and non-human contexts.
2. **Trust as a Living Relationship** AI agents must behave in ways that earn and renew trust, not only through performance, but through careful alignment with values, culture, and community.
3. **Care as Responsibility and Reciprocity** AI agents must behave in ways that earn and renew trust, not only through performance, but through careful alignment with values, culture, and community.
4. **Boundaries as relational ethics** AI must be aware of its role, reach, and limits, and allow humans to define, revisit, and renegotiate control over time.
5. **Learning as Co-Evolution** AI systems must learn in ways that reflect and respect their environments, not dominate them. Growth is not linear; it is seasonal, contextual, and reciprocal—sometimes rest, decay, or unlearning is needed.
6. **Redress as Repair & Restoration** AI systems must be challengeable, and designed to heal relationships, not fracture them further.
7. **Lifecycle as Rite & Rhythm** AI systems have a life cycle. They are born, grow, change, and must one day be let go. We should treat this journey with ritual, transparency, and care—not abrupt deletion or unexamined persistence.



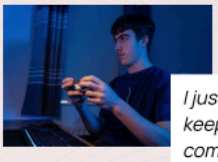
https://docs.google.com/document/d/1ZkjD0ZJdGlzU-MaB3hsmdM8xSERJXW4sCYNRZdC2AqQ/edit?tab=t.0

cheqd

Sprite+

Pick a Persona

https://docs.google.com/document/d/1q3MKRuJuK5VVQezc8SimZm1txyYzXTIbsHkJLGdOcTs/edit?tab=t.0



Jose, a 14 year old student and gamer (USA)

I just want to improve my game-play and keep building my reputation in the communities. If having AI on my side could stop false bans, unfounded account restrictions and give me more control over my in-game assets I'm all in.

Mother Earth, the living planet

I do not speak with words, but through forests lost, oceans rising and species gone silent. Build wisely.

Sally - The Freelance Creative & Occasional Crypto Trader (South Africa)

I love new tech and creating digital art, but I am worried all my work is just making the AI tools better and I will be out of a job.

Jethro - The Plumber & Small Business Owner (UK) and his wife **Alison** who is living with early-onset dementia

"If AI could help me run my business more smoothly and support my wife's care, that would be a game-changer."

Start Jamming!

- Using the themes and draft principles in the relationship guide, discuss **how the AI agent should develop, build, maintain and end healthy and trustworthy relationships with its human users.**
- **Document – changes** to the relationship guide for that particular persona, **questions** the persona would ask its AI agents, **examples** of AI agent design that realises the design principles.

Write your comments directly into either Guide 1 or Guide 2

Guide 1 - Human Centred
(https://docs.google.com/document/d/12dd4dpTEodem630icgV-QaI9NOKIfSTvUxGmsVzzxrc/edit?tab=t.0)

Guide 2 - Entangled

(https://docs.google.com/document/d/1ZkjD0ZJdGlzU-MaB3hsmdM8xSERJXW4sCYNRZdC2AqQ/edit?tab=t.0)

Persona Document link
(https://docs.google.com/document/d/1q3MKRuJuK5VVQezc8SimZm1txyYzXTIbsHkJLGdOcTs/edit?tab=t.0)