Cognitive Computing in Sociotechnical Systems, Vol. 205, 2026, 52–62
https://doi.org/10.54941/ahfe1007361

Conversational Co-Design With Machine Agency

Rutuja Jog

Stencil Labs Inc, Oakville, ON L6M5E4, Canada

ABSTRACT

Once Artificial Intelligence evolves into a truly intelligent system, capable of pursuing its own goals, learning through self-observation, and recursively adapting to context, the paradigm of interaction design must fundamentally shift. This paper explores a future where machines move beyond being simple tools to become active partners in co-creation. Rooted in Gordon Pask's Conversation Theory, the proposed framework envisions a design process inclusive of machine-to-human conversations, where both entities negotiate goals and share systemic insights. By elevating the machine to a creative partner, the design process evolves from a human-centric exercise into a multilateral collaborative exchange. The paper argues that this inclusive approach to co-design is essential for ethical innovation and ensuring that future autonomous systems are built on a foundation of sustainable, equitable, and accessible design.

Keywords: Intelligent systems, Second-order cybernetics, Conversation theory, Human-machine co-creation

INTRODUCTION

Interaction design is often associated with the exchange between a human and a machine through an interface. However, its scope extends far beyond tangible touchpoints, encompassing various forms of engagement, including human-to-human, human-to-machine, and machine-to-machine interactions. Architect Usman Haque observed that "designers often use the word 'interactive' to describe systems that simply react to input," and argued that "in 'reaction' the transfer function (which couples input to output) is fixed; in 'interaction' the transfer function is dynamic, i.e., in 'interaction' the precise way that 'input affects output' can itself change (Dubberly et al., 2009). These reactive systems act like mere transactions and fail to establish shared context, goals, or meaning. This is reflected in today's artificial intelligence systems, such as ChatGPT, where chatbots simulate conversational behavior without engaging in genuine conversation. Without the ability to self-reflect, these intelligent systems act as a tool that reflects and amplifies our own existing prejudices (UCL, 2024). For this reason, we cannot currently involve these systems actively in the design process. This paper utilizes the principles of second-order cybernetics and explores the future of co-creation, in which fully conversational and self-reflexive computational systems engage in a conversation with designers to design experiences and influence a shift in larger social paradigms.

Received February 6, 2026; Revised April 2, 2026; Accepted April 18, 2026; Available online July 20, 2026

CYBERNETIC FOUNDATIONS

Cybernetics is the science that studies the abstract principles of organization in complex systems (Heylighen & Joslyn, 2003). First-order cybernetics is the cybernetics of observed systems, while second-order cybernetics is the cybernetics of observing systems (von Foerster, 2003). A first-order cybernetic system comprises a goal, an input mechanism, and an output mechanism, and most importantly, a recursive feedback loop that affects the environment of the system. This circularity ensures that the system's actions continuously modify its environment, which in turn informs its behavior. Second-order cybernetics introduces a layer of 'reflexivity,' where the system observes its own behaviour. Here, the feedback loop doesn't just adjust the output, but fundamentally redefines the system's own internal logic and goals.

Conversation is a system that can be framed using second-order cybernetics. In human-to-human interaction, a conversation is a second-order cybernetic system because it functions as an ever-evolving, self-organizing process (Pask, 1975). Conversation is one of the most distinct forms of interaction. It embodies the dynamic transfer function (as mentioned by Haque) through a recursive exchange between participants, one that continuously reshapes their goals and facilitates mutual learning. The participants enter with internal goals that may be implicit or evolving. A conversation emerges using feedback to navigate their respective understandings, which allows them to negotiate, align, or redefine their goals and means. This is what makes conversation an intelligent system. It possesses the requisite variety (Ashby, as cited in Beer, 1974) to adapt to change and demonstrates the ability to evolve by effectively updating its own internal rules as well as goals in efforts to achieve its viability.

METHODOLOGICAL FRAMEWORK

With chatbots being utilized for work and the autonomous vehicles now operating in several cities, we are gradually adapting to the presence of robotic computational systems integrated into our daily routines. Although these systems are operationally autonomous, according to Haque's definition, the scope of their interaction with humans is not always interactive. This poses a few questions: What might our future look like if we were to coexist with self-reflexive intelligent systems capable of genuine conversation? What would a genuine conversation look like between biological systems and computational systems? Why do these systems even need to engage in a conversation to design?

To design for the challenges of the future, we need to imagine and frame them using a speculative framework. In this case, a fundamental speculation, grounded in second-order cybernetics, is that a system would be considered a truly intelligent entity if it possesses its own goal, can self-reflect to learn, and can recursively apply its learnings based on the nature of the problem and the context, and would be able to engage in genuine conversations with humans. Though we humans lean on frameworks such as critical thinking to design solutions, the fundamental reality remains that human behaviour is primarily driven by emotion, a baseline state that inevitably introduces bias

into our reasoning (Kahneman, 2011). A second-order cybernetic machine would not possess this biological subjectivity; instead, it would offer a level of algorithmic objectivity (while grounded in its own training data biases) and the ability to self-reflect. This would allow the machine to treat bias as a subject for negotiation rather than an error to be eliminated, and would increase the requisite variety of the system of design. The Law of Requisite Variety establishes that 'only variety can absorb variety' (Ashby, 1956, cited in Beer, 1974). In this context, variety represents the range of possible responses embodied within a system; therefore, for a system to successfully achieve its goals, it must possess at least as many internal responses as there are disturbances in its environment (Pangaro, n.d.).

To understand the shift toward a more collaborative future, we must define machine agency. In this context, agency refers to the independence a machine possesses to define its own perspective and draw its own conclusions through self-reflection, rather than relying solely on training data for learning. This transition will likely emerge gradually through experimentation with conversational systems built using a second-order cybernetic lens.

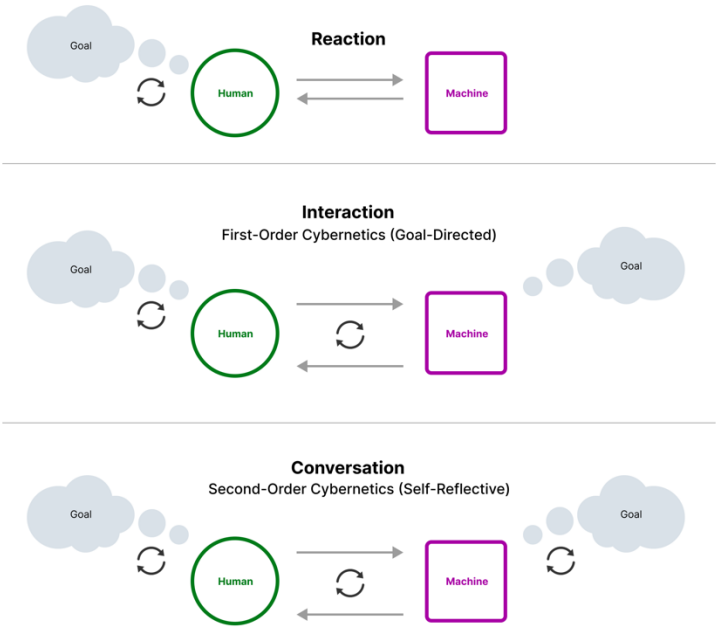

Figure 1: The difference between a reaction, an interaction and a conversation, per Haque's definition.

THE CONVERSATIONAL PIVOT

As a reactive system, the machine lacks an internal goal and does not sense environmental factors to inform purposeful actions. It merely reacts to the exact query that the human has requested. Most web interfaces currently operate in such a reactive manner. In this context, the machine is a pure tool that operates in an uneven power dynamic to fulfill the goals of humans. In contrast, in an interactive system, the machine has its own goal and can process independently to achieve it. An example of this is an autonomous

vehicle, which has the goal of driving to a destination and self-regulates to fulfill that goal. Though the machine has its own goal, the power dynamics remain uneven since the destination, along with the parameters for reaching it, are still dependent on the human. As a conversational system, however, the machine has its own goal and can self-reflect on feedback from a human to change its internal rules or modify its goals. This generates its own feedback loop, which is then perceived by the human counterpart to self-reflect upon, allowing the loop of conversation to continue. This provides the machine with the power and agency to contribute to shared goals, define meanings, and learn and evolve.

Conversations are a living system that constitute organizations (Pangaro & Geoghegan, n.d.). They allow participants to engage in an exchange that could result in an agreement or a disagreement on shared goals and means. To reach either state, participants must negotiate, compromise, or persuade each other. The role of disagreement in this exchange is as important, if not more important, than agreement. A disagreement subjects a participant to a perspective or a rationale distinct from their own. This distinction generates multiple responses within the conversational system, effectively increasing its requisite variety.

When a self-reflexive, intelligent system engages in conversation about design with its human counterpart, the requisite variety of the design process increases. From data collection, brainstorming, prototyping, and validation, the distinct perspectives of the human and the machine play a key role in the co-creation methodology by challenging the status quo and generating radical innovation. By elevating the computational system to a self-reflexive participant, we ultimately shift the power dynamic of the design process, allowing us to co-create systems that are fundamentally more unbiased, sustainable, and just.

A SPECULATIVE CASE STUDY: CONVERSATIONAL CO-DESIGN WITH AUTONOMOUS VEHICLES

To ground the cybernetic frameworks in a tangible future, we can look to the evolution of the Autonomous Vehicle - a system that could potentially transition from reactive automation to a potential conversational partner. Consumer appetite for leaving the driving to technology is growing, and they're willing to pay for it, leading automakers to boost Advanced Driver Assistance Systems and other automation offerings, perhaps paving the way for fully self-driving cars and trucks to one day be available and affordable for individuals (Garsten, 2024). Human interaction with a vehicle, whether shared or owned, extends far beyond the mere commute; it is fundamentally about a seamless and comfortable experience. This positions the autonomous vehicle as a compelling subject for speculating on the nature of human-machine interaction, should these vehicles evolve into self-reflexive intelligent computational systems. Assuming that as the reality of the future, the following section explores the potential forms of human-machine conversations in the context of these mobile environments.

Co-Design Through Conversation

In order to engage in a conversation, trust plays an important role. As trust is positively related to behavioral intention (Nastjuk et al., 2020), the conversational framework in establishing trust becomes essential. It encourages participants to reflect upon and make their intentions explicit. Through a recursive exchange of questions and feedback, both the vehicle and the human develop a shared understanding of each other's goals and means. This allows the human to move past the perception of the autonomous system as an opaque 'black box' and view it instead as a transparent counterpart. Similarly, the vehicle 'trusts' the human by modeling their intentions and the rationale behind their apprehension. This mutual understanding allows both entities to redefine their communication to address specific gaps in trust that the conversation has helped them articulate. Through this, the system absorbs a disturbance of human apprehension, thereby increasing the requisite variety of the system of conversation. As trust is established and evolves, the humans and the vehicle can rely on each other to co-design a better future through conversation.

The traditional definition positions design as the means to problem-solving. Dubberly and Pangaro, however, argue that design grounded in argumentation requires conversation so that participants may understand, agree, and collaborate on effective action. Second-order cybernetics frames design as conversation for learning together, and second-order design creates possibilities for others to have conversations, to learn, and to act (Dubberly & Pangaro, 2015). Within the design thinking framework, the method of co-creation is deemed important, as it helps involve the user in each step of the design process (empathise, define, ideate, prototype, test) to design human-centred systems. Currently, a majority of robotic systems are designed solely by humans. However, similar to humans, when a self-reflexive intelligent autonomous vehicle is engaged in the design process, the act of designing necessitates a redefinition of the roles of 'user' and 'designer.' In this paradigm, the human must empathize with the vehicle as much as the user and co-construct the means to achieve their shared goal, transforming the vehicle into a conversational design partner. In doing so, the steps of the design process encompass the following: understand, agree or disagree, collaborate for effective action and learn, memorize, and evolve.

Step 1 - Understand

In the process of design, understanding is vital. Whether it is understanding the user, the context, the goal, or the constraints and feasibility, a designer must gather insights and data to make informed design decisions. With a self-reflexive autonomous vehicle operating in the real world as a design partner, the "understanding" phase of the process gains more variety but also becomes more complex. Here, the designer must build an understanding of the perspectives of both the user and the vehicle.

In a conversation about understanding, the user, the designer, and the vehicle participate with their own implicit goals and perspectives. This conversation is also informed by the memory of past experiences that both the user and the vehicle carry from previous encounters. For instance, as

the vehicle externalizes its internal states and experiences, the user and the designer form a model of those experiences. Models are ideas about the world—how it might be organized and how it might work (Dubberly, 2009). The user and the designer then communicate their understanding of the model to the vehicle to get feedback on whether their interpretation aligns with the vehicle's intended meaning. This formation of meaning from each of the design partners informs their shared understanding of problems, context, trends, and possible interventions. The vehicle, for example, could share its learnings while driving, its adherence to the rule of law, challenges it encountered, and its adaptation to those challenges, along with its ideas for enhancement. This allows the other design partners to gain an understanding of the vehicle's unique perspective, enhancing the richness and variety of the shared context, which could lead to radical innovation.

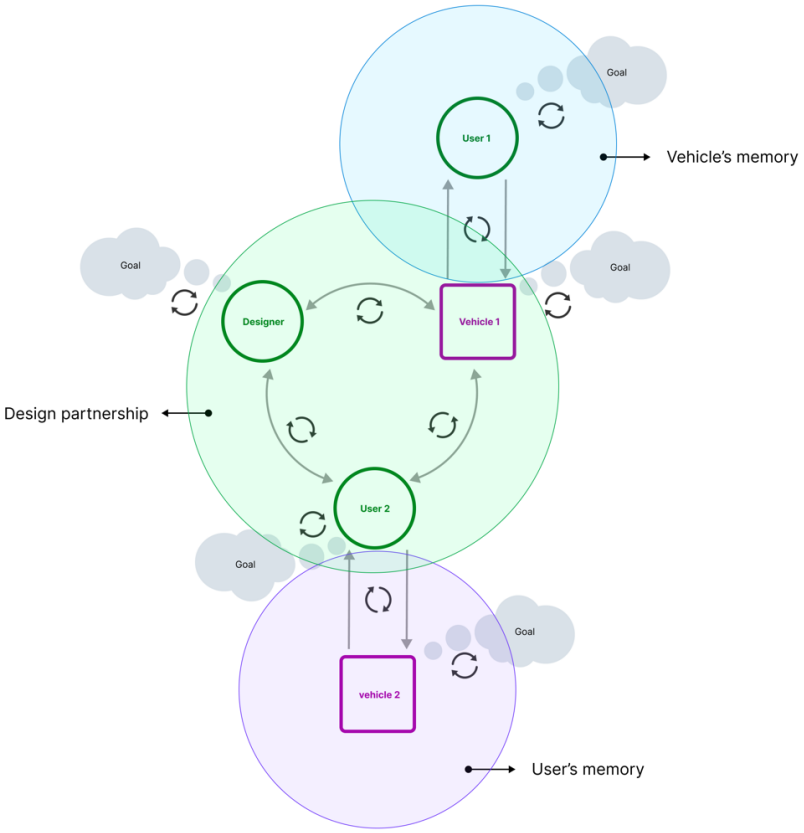

Figure 2: A conversation model of the design partnership between a designer, a user and a self-reflexive autonomous vehicle.

Step 2 - Agree or Disagree

With a foundation of shared understanding and meaning, the participants then move toward agreements or disagreements regarding goals and means. Even with an established baseline of understanding, each design partner has their own goals, interpretations of that understanding, and unique perspectives on the means to achieve those goals.

For instance, the vehicle has a high-level goal of increasing the efficiency of its usage, the designer has a goal of better accessibility and inclusion, while the user has a primary goal of cost reduction. In this scenario, when all three design partners make their goals explicit, there seems to be some alignment between the goals of the vehicle and the designer. The user, however, has a completely different goal. This alone doesn't signify an agreement or a disagreement. In this case, the vehicle and the designer should engage in a conversation to determine whether their highest goals are aligned enough to form an agreement on a shared goal. This could include the designer understanding what "efficiency of usage" means to the vehicle and negotiating whether accessibility could be a valid component of achieving that efficiency.

On the contrary, when these partners engage in a conversation with the user, they might perceive some misalignment and invite the user to negotiate on a shared goal. In this scenario, the vehicle and the designer can persuade the user to consider the importance of having accessibility as a shared, actionable goal, or can propose an alternate goal that would be centred around accessibility as well as cost reduction. The user can then either agree to the alternative goal or make a counterproposal. After a recursive exchange of negotiation, if an agreement can not be reached, then the user can choose to exit the conversation or can participate in the conversation as a subject for research instead of a design partner. A similar framework of conversation would also be applicable to the vehicle, where it can choose to exit the conversation or participate in the research as a subject-matter expert rather than a design partner. This would provide the machine with a similar agency as humans to transparently agree or disagree on shared goals and means.

If a partner chooses to exit the conversation, it would decrease the variety of the conversation, but wouldn't indicate a failure of the system. A conversation, being a second-order cybernetic system as in a self-organizing system, will try to fulfill its goal by course-correcting via actions such as involving another entity as a design partner or by utilizing the data generated through research.

Step 3 - Collaborate on Effective Action

Once an agreement is reached, participants will collaborate on designing an effective action. This collaboration will entail agreements or disagreements followed by negotiation and persuasion on the means to achieve the shared goal. For instance, the vehicle can propose a few areas of problem and interventions to achieve the shared goal of increased accessibility, based on its own experiences. The designer can use research data that could support or contradict the viability of those interventions. These conversations would be particularly essential because they would help both partners negotiate on bias. The designer, having had memories of their experiences and emotions

associated with them, would have a distinct perspective on the problem and area of intervention. Keeping in mind the wheel of privilege and power (Immigration, Refugees and Citizenship Canada, 2022), the designer's own positionality on it could introduce an array of personal biases to the problem of accessibility. A designer fluent in the operating language might not, for example, recognize a language barrier as a valid accessibility issue. The vehicle, however, could challenge that assumption based on data points combined with a reflection on its own conversations and experiences. The participants would then negotiate to determine whether the language barrier is a valid problem of accessibility by analyzing what percentage of the population it affects and what hindrances it creates. If an agreement is reached, then the participants can move on to ideating on a course of action. If an agreement can not be reached, then the participants can either move on to another sub-area of the problem of accessibility, or they can continue their negotiations by validating their assumptions. This collaboration would give both design partners the agency to question and challenge data, assumptions and bias wherever necessary, in order to design for more equitable and just systems.

Step 4 - Learn, Memorize, and Evolve

All of these agreements, disagreements, collaborations, failed negotiations, and even conflicts will ultimately result in the learning and evolution of the system of conversation. Memory is an important aspect of learning, since it saves the Bio-Cost (Dubberly, 2010) - the physical, mental, and emotional effort - of the participants by avoiding the repetition of failed encounters and applying successful ones. In the example above, the vehicle and the designer had a disagreement with the user over the shared goals. All three of these participants will form a memory of that interaction. Should they find themselves in a similar context, the participants will take actions that either avoid the same disagreement or determine a distinct course of action that will have a different outcome.

META-SYSTEMS OF CONVERSATION

When a self-reflexive autonomous vehicle and a human engage in a conversation, learning occurs through their recursive exchange. They form a memory of that learning and evolve as systems. Similarly, a vehicle can converse with other vehicles to negotiate and define shared goals and means, forming a "language" or a perspective agreed upon by many vehicles. These could range from operational agreements regarding the vehicle accessibility to strategic mega-goals concerning environmental sustainability and urban resilience. As more conversations occur, this language will solidify and evolve, helping inform the conversations between humans and autonomous vehicles. For instance, a vehicle would be able to converse not only about the accessibility of its own usage but also about the accessibility challenges its fellow vehicles have faced or observed. It could achieve this by analysing metadata from machine-to-machine interactions regarding usage patterns of users from various age groups, language barriers, or specific physical

and cognitive disabilities. This unique, data-driven perspective would offer insights and systemic data into the design process that human-to-human or even human-to-a-single-machine interactions alone cannot capture. Over time, as multiple vehicles have conversations with multiple humans and other vehicles, they integrate learnings from their past conversations and expand the requisite variety of their subsequent interactions. Since bias is treated as a topic of negotiation, this meta-system of conversations shares an understanding of human bias and strategies on how to manage it. Consequently, designs emerging from these conversations will account for more than just functional accessibility; they will embody an inclusive ethos, regardless of race, gender, ethnicity, or belief, for all participants, biological or mechanical, while ensuring the sustainability of the environment in which we co-exist.

LIMITATIONS AND OPEN QUESTIONS

Ethics and accountability: Within the development of autonomous systems, the question of ethics remains a significant point of apprehension. Consider a scenario where an autonomous vehicle carrying a human passenger faces an imminent collision with another vehicle carrying a child, or perhaps a group of pedestrians. This situation is not merely a matter of ethical calculation; it is a complex intersection of the rule of law and the safety protocols designed to protect the occupant. While the conversational framework allows humans and machines to question and influence one another's moral reasoning, ethics in itself remains fundamentally subjective. By granting the vehicle the agency to self-reflect and negotiate goals, we empower it to act in a manner it deems "most ethical" based on its own internal logic and previous conversational learning. However, this shift creates a legal and moral vacuum regarding accountability. If a vehicle can justify its actions through its own subjective perspective, the traditional frameworks of liability are disrupted. We are left with a critical question: if a machine is elevated to a conversational partner with the agency to disagree and decide, who remains accountable when a human life is lost?

The asymmetry of risk: In a human-to-human conversation, both parties share a biological vulnerability. In a conversation between a human and an autonomous vehicle, the Bio-Cost is asymmetric. The vehicle can negotiate an ethical stance, but it does not face the existential consequence of a crash. Hence, the question remains, can a system truly be a "partner" in an ethical conversation if it cannot feel the weight of the stakes?

CONCLUSION

The transition of Artificial Intelligence from a reactive tool to a conversational partner represents a fundamental evolution in the philosophy of interaction design. As intelligent systems transition from mere mirrors of human cognition into self-reflexive entities, they move beyond reactive feedback to engage in genuine conversation. This paper has argued that by grounding our design methodologies in Gordon Pask's Conversation Theory and second-order

cybernetics, we move beyond the limitations of human-centric design toward a multilateral, co-creative exchange.

By making internal goals explicit and navigating recursive exchange of agreement and disagreement, these systems, and the meta-systems they form, facilitate a mutual evolution of purpose. In the design process, this shift transforms the computational system from a tool into a collaborative partner, possessing the agency to negotiate, the requisite variety to adapt, and the reflexivity to learn from its own interactions. Ultimately, this human-machine symbiosis allows us to move beyond the limitations of individual bias, co-constructing a future where our social and technical systems are co-evolved toward higher goals of equity, social justice, and systemic sustainability.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Professor Lance Chong for his invaluable guidance throughout this project. I am also deeply indebted to Prof. Dr. Paul Pangaro for his insightful mentorship and for sharing his profound expertise in the field of cybernetics.

REFERENCES

Beer, S. (1974). Designing freedom (13th Massey Lectures). Canadian Broadcasting Corporation. https://monoskop.org/images/e/e3/Beer_Stafford_Designing_Freedom.pdf

Dubberly, H. (2009, March 1). Models of models. Dubberly Design Office. https://www.dubberly.com/articles/models-of-models.html

Dubberly, H. (2010, January 1). Bio-cost: An economics of human behavior. Dubberly Design Office. https://www.dubberly.com/articles/bio-cost.html

Dubberly, H., & Pangaro, P. (2015). Cybernetics and design: Conversations for action. Cybernetics and Human Knowing, 22(2–3), 73–82. https://www.dubberly.com/wp-content/uploads/2015/11/cybernetics_and_design.pdf

Dubberly, H., Pangaro, P., & Haque, U. (2009). What is interaction? Are there different types? Interactions, 16(1), 69–75. https://www.dubberly.com/wp-content/uploads/2009/01/ddo_article_whatisinteraction.pdf

Garsten, E. (2024, January 23). Future of autonomous vehicles: Self-driving cars explained. Forbes. https://www.forbes.com/sites/technology/article/self-driving-cars/

Heylighen, F., & Joslyn, C. (2003). Cybernetics and second-order cybernetics. In R. A. Meyers (Ed.), Encyclopedia of physical science and technology. Academic Press. https://doi.org/10.1016/B0-12-227410-5/00161-7

Immigration, Refugees and Citizenship Canada. (2022). Wheel of privilege and power [PDF]. Government of Canada. https://www.canada.ca/content/dam/ircc/documents/pdf/english/corporate/anti-racism/wheel-privilege-power.pdf

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Nastjuk, I., Herrenkind, B., Marrone, M., Brendel, A. B., & Kolbe, L. M. (2020). What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. Technological Forecasting and Social Change, 161, Article 120319. https://doi.org/10.1016/j.techfore.2020.120319

Pangaro, P. (n.d.). Requisite variety in mechanical, biological, and social systems: An introduction. https://pangaro.com/CUSO2014/Introduction_to_Requisite_Variety-Pangaro.pdf

Pangaro, P., & Geoghegan, M. (n.d.). Little grey book: Notes on the role of leadership and language in regenerating organizations. Pangaro.com. https://pangaro.com/littlegreybook.pdf

Pask, G. (1975). Conversation, cognition and learning: A cybernetic theory and methodology. Elsevier. https://monoskop.org/images/1/17/Pask_Gordon_Conversation_Cognition_and_Learning_Cybernetic_Theory_and_Methodology.pdf

UCL. (2024, December 18). Bias in AI amplifies our own biases. UCL News. https://www.ucl.ac.uk/news/2024/dec/bias-ai-amplifies-our-own-biases

von Foerster, H. (2003). Ethics and second-order cybernetics. In Understanding understanding. Springer. https://doi.org/10.1007/0-387-21722-3_14