Designing the interaction between humans and autonomous systems: the role of behavioral science

Van Rooy, Dirk

Antwerp Centre for Responsible AI (ACRAI), University of Antwerp, Antwerp, Belgium
dirk.vanrooy@uantwerpen.be

doi.org/10.21606/iasdr.2023.457

In recent years, the use of autonomous systems has grown rapidly in both the industrial and military sectors. These systems have the potential to revolutionize the way we live and work, from self-driving cars and drones to automated factories and military equipment. However, it has been argued, for this to happen, researchers and designers need to (re)consider the interaction between humans and machines, as it plays a significant role in ensuring the safe and efficient operation of autonomous systems (De Regt & Gagnon, 2020; Janssen et al., 2019). In this paper, I will argue that a closer integration with applied behavioural science could be beneficial for both researchers and practitioners involved in the design of the Human–machine interaction. To that effect, I will discuss a number of insights from behavioural science and how they could inform the design of human-machine interaction.

Keywords: human-machine interaction; behavioural science; autonomous systems; design for interaction

1 Introduction

As it emerged in the late 19th and early 20th centuries, and researchers began to study human behavior and cognition in more systematic ways, behavioral science as a general discipline has had an increasing impact on design (Brown, 2008; Cash et al., 2022). A classic, early example is that of Gestalt psychologist Max Wertheimer, who conducted groundbreaking research on visual perception in the early 20th century that highlighted the importance of considering how people perceive and interpret visual stimuli. His work on the "phi phenomenon" for instance, revealing how people perceive motion in static images, had profound implications for the design of visual displays. Similarly, the work on operant conditioning and behaviorism in the mid-20th century demonstrated how human behavior could be shaped and modified through reinforcement and punishment, which has had significant implications for designing persuasive systems that seek to encourage desired behaviors (and discourage undesirable ones) (Deterding, 2012; Wenker, 2022). Donald Broadbent's research, which demonstrated how people selectively attend to information based on its relevance and importance, has been quite influential in the design of user interfaces that prioritize relevant information and minimize distractions, such as in the design of notification systems on smartphones and other devices (Broadbent, 1990). And most designers would be familiar with the concept of mental models, which reflect the understanding people form of how things work, which has proven useful in creating interfaces that are user-friendly and efficient (Norman, 2013).

2 Behavioral science and human-machine interaction

In this paper, I will focus on the application of behavioral science to the interaction between humans and autonomous systems. Automation typically refers to the use of technology, machinery, or computer systems to perform tasks or processes with minimal human intervention. Autonomous systems are a subset of automation, representing a higher level of sophistication and independence, and are typically defined as self-governing, self-regulating or self-operating entities (Lyons et al., 2021). Importantly, they can perform tasks or make decisions without continuous external control or direct human intervention, and are thus designed to operate with a certain degree of independence. Going forward, "machine" in the context of this paper refers to an autonomous system that can operate and make decisions in a partially or fully autonomous manner. It can be argued that the increased autonomy of machines has shifted the human-machine relationship from mere interaction towards cooperation and collaboration (Lyons et al., 2021; Schelble et al., 2020). In this context, it has become even more important to have insight not just in human decision making, but also into how humans understand the behavior of the machine (algorithm), and whether or not they accept and trust the machine (Lyons et al., 2021; Xiong et al., 2022). Behavioral science investigates this type of understanding extensively through the exploration of key factors such as mental models (Mathieu et al., 2000) and situation awareness (Endsley, 1995), which has provided significant insights into how human operators comprehend the surrounding environment, including the machine, while also shedding light on the human decision-maker's understanding of tasks, particularly in dynamic contexts.

So perhaps not surprisingly, behavioral science has already played a crucial role in humans and autonomous systems design across various domains (Cross & Ramsey, 2021). In the context of autonomous vehicles, it has shed light on the "out-of-the-loop" problem, where excessive information and feedback can overwhelm drivers and reduce performance (Endsley & Kiris, 1995). This has informed the design of autonomous vehicle interfaces and smart home systems (Choi & Ji, 2015; Lee et al., 2015), ensuring that user interactions are optimized to avoid overload and foster a more effective human-machine partnership (Strengers, 2013). AI-powered chatbots and virtual assistants heavily rely on understanding user intent and providing accurate responses to create a positive user experience. Thanks to research in human language processing and conversation dynamics, chatbot interfaces have been refined to improve user engagement and communication (Kuhail et al., 2023). Behavioral science research on automation's impact on trust, situational awareness, and accountability has been instrumental in developing interactions that support cognitive load management, decision-making, and effective communication with autonomous systems (Endsley & Kiris, 1995; Norris, 2018; Sparrow, 2009), particularly in military settings where human operators of autonomous systems often face challenges with workload management and decision-making (Bewley et al., 2014).

3 Behavioral economics

Behavioral science has thus already significantly contributed to human-machine interaction (HMI), and overall we are seeing an increasing focus on the cognitive and behavioral factors that influence human interaction with autonomous systems (Hopko et al., 2022; Krausman et al., 2022). However, this research can be overwhelming: Behavioral science research is often complex and may not provide clear and actionable guidelines for design practitioners. Translating abstract behavioral concepts into practical design solutions can be challenging, particularly as design practitioners may also have limited knowledge or awareness of behavioral science principles and research, and may not be familiar with the latest findings. Applying behavioral science insights often requires interdisciplinary collaboration between designers, psychologists, and other experts in the field, and design practitioners may face challenges in collaborating with other experts when trying to integrate behavioral science perspectives into their design processes. Where to begin as a design practitioner when one wants to apply behavioral scientific insights to the design of human-systems interactions? A number of researchers have argued that a good place to start is some of the basic literature in behavioral economics, as it tends to be more practical and applied, and therefore easier to adapt than a lot of core psychological research (see for instance Voyer, 2015). The field of behavioral economics has emerged as a prominent area of study with significant implications for design, as it seeks to understand how people make decisions in real-world situations and how these decisions can be influenced by cognitive biases and social norms. It deals specifically with decision making under uncertainty, which makes it particularly useful for our purpose: By understanding the cognitive biases that influence decision-making and by developing effective strategies to mitigate them, it can potentially help designers create interactions with autonomous systems that ensure that humans understand the logic and decision-making processes of complex algorithms, while also mitigating the impact of social and cognitive biases (see for instance Bertrand et al., 2022; Xiong et al., 2022).

3.1 Fast and slow thinking

One of the key distinctions made in behavioral economics research is that between two types of thinking: Type I and Type II (Kahneman, 2013) (see Table 1). Type I thinking, also known as intuitive or automatic thinking, is fast, unconscious, and relies on heuristics or mental shortcuts. It is often influenced by cognitive biases, such as the availability bias or the confirmation bias, which can lead to deviations from rational decision-making (see Table 2 for examples of cognitive biases). On the other hand, Type II thinking, also known as reflective or deliberative thinking, is slow, conscious, and analytical, involving careful consideration of options and consequences. It involves conscious cognitive processing that requires effort, focus, and attention, and is a slow and deliberate process that is often used for complex tasks, problem-solving, and decision-making. Type II thinking allows individuals to critically analyze information, consider different options, and make informed decisions based on careful evaluation of available evidence. However, Type II thinking can be easily overwhelmed in high-stress situations or when individuals are required to manage multiple tasks simultaneously (AlKhars et al., 2019; Nurse et al., 2022).

3.2 Cognitive biases

Research into cognitive biases, and their impact on Type I thinking, is already impacting design practices and decisions. For example, the anchoring bias refers to the tendency of individuals to rely too heavily on the first piece of information encountered when making decisions, even if it is arbitrary or irrelevant (Meppelink et al., 2019). We can see this bias being put to work on e-commerce websites, where the initial price presented to users acts as an anchor that shapes their perception of value and affects their willingness to pay. The framing effect refers to the phenomenon where the way information is presented or framed can influence decision-making outcomes (Stea & Pickering, 2019).

Table 1. Characteristics of system 1 and system 2.

System 1	System 2
Does not require working memory	Requires working memory
Automatic	Controlled
Fast	Slow
High capacity	Limited capacity
Nonconscious	Conscious
Independent of cognitive ability	Correlated with cognitive ability

Table 2. Cognitive biases affecting System 1 thinking.

Cognitive bias	Description
Confirmation bias	To look for or to interpret evidence to support prior hypothesis rather than look for disconfirming evidence.
Anchoring effect	To rely heavily on one piece of information when making decisions (usually the first piece of information acquired: the 'anchor').
Availability bias	Judgments of likelihood or percentages based on ease of recall (greater 'availability' in memory) rather than on actual probabilities.
Framing effect	To draw different conclusions from the same information, depending on how that information is presented.
Loss aversion	To view losses as looming larger than corresponding gains.
Sunken-cost fallacy	To allow previously spent time, money, or effort to influence present or future decisions.
Social proof	Also often referred to as the Bandwagon effect. To do (or believe) things because many other people do (or believe) the same.

For example, a study by Tversky and Kahneman (1981) showed that people tend to be risk-averse when decisions are framed in terms of gains (e.g., "you have a 70% chance of winning $100") but risk-seeking when decisions are framed in terms of losses (e.g., "you have a 70% chance of losing $100"). This insight has already informed the design of products and systems where decisions involve risks, such as financial investments or medical treatments (Traut, 2023). The concept of social proof suggests that people tend to conform to the actions of others in uncertain or ambiguous situations (Cialdini & Jacobson, 2021), which has informed the design of products or systems that rely on user-generated content, such as online reviews or ratings. By highlighting the actions of others or showcasing the popularity of certain choices, social proof is leveraged to influence users' decisions and encourage desirable behaviors, such as purchasing a product or signing up for a service. Designers have also been leveraging the power of "defaults", which refer to the pre-set options or choices that are presented to users that can significantly influence decision outcomes. For example, a study by Johnson and Goldstein (2003) found that changing the default option for organ donation from opt-in (where individuals have to actively choose to be a donor) to opt-out (where individuals are automatically considered donors unless they actively choose not to be) can dramatically increase the number of organ donors. This insight has been applied in various design contexts, such as in online form design, where designers can strategically set default options, or present options in particular ways (so-called choice architectures) to encourage certain behaviors or choices (Mertens et al., 2022).

3.3 Enhancing human-machine interaction

HMI design has traditionally focused on usability principles, such as efficiency, effectiveness, and learnability. Behavioral scientific insights have already, and continue to have, a significant impact in that domain (Effie Lai-Chong Law, Ebba Thora Hvannberg, Gilbert Cockton, n.d.; Ferreira et al., 2020; Jeffries & Wixon, 2007; Zaharias & Poulymenakou, 2006). Some work has begun to explore the impact of cognitive biases on the interpretation of AI models (Kliegr et al., 2021), however it can be argued that the further integration of behavioral economics principles, and in particular a better understanding of the interplay between Type I and Type II thinking, can be beneficial for HMI design (Bertrand et al., 2022; Xiong et al., 2022). In this section, I formulate a number of key insights and, where possible, formulate some initial design guidelines (see Table 3 for an overview).

3.3.1 Mitigating cognitive biases

The operation of autonomous systems in complex and uncertain environments often necessitates rapid decision-making, a context in which human operators tend to rely on Type I thinking (Mayer, 2014). However, Type I thinking, which is characterized by intuitive and heuristic-based decision-making, has been shown to be susceptible to numerous biases and errors. In the context of human interactions with autonomous systems, confirmation bias can lead to inaccurate or suboptimal decisions when operators selectively focus on information that aligns with their preconceived notions about the system's capabilities or performance, while disregarding contradictory information. For example, an operator of an autonomous vehicle may rely on information that supports the notion that the vehicle is performing optimally, despite receiving warning signals indicating a malfunction. An anchoring bias can lead to inaccurate or suboptimal decisions when operators place disproportionate weight on initial information or data provided by the system, without thoroughly evaluating additional information or considering alternative options. For example, an operator of an unmanned aerial vehicle may anchor on the initial altitude information provided by the system, without cross-checking it with other sources, leading to a wrong decision about the vehicle's position in the airspace.

Research has already demonstrated how insights into cognitive biases play a pivotal role in shaping human-machine interaction. Specifically, research in the field of eXplainable Artificial Intelligence (XAI) has demonstrated its potential in addressing various challenges in algorithmic decision-making, particularly in countering framing effects and confirmation bias (Danry et al., 2020; Springer & Whittaker, 2019; Wang et al., 2019). For instance, providing explanations like local feature importance and presenting explanations gradually or upon request to avoid contradicting users' expectations has been shown to mitigate a range of cognitive biases (Springer & Whittaker, 2019). Wang et al. (2019)

delve into the significance of representativeness and availability bias in the context of medical diagnosis and propose countermeasures such as presenting prior probabilities and outcome prototypes to alleviate these biases. It has also been shown that incorporating arguments for non-predicted outcomes (Wang et al., 2019; Weld and Bansal, 2019; Bussone et al., 2015) helps users comprehend an AI's reasoning and build trust in its predictions. Employing strategies such as delaying the presentation of AI predictions and explanations (Wang et al., 2019; Weld and Bansal, 2019; Bussone et al., 2015; Lai and Tan, 2019) can foster more thoughtful and reflective interactions. Furthermore, integrating cognitive forcing functions (Buçinca et al., 2021) enhances the decision-making process by nudging users to consider alternative perspectives and avoid hasty judgments. Lastly, providing uncertainty estimates (Wang et al., 2019) empowers users to gauge the reliability of AI predictions, promoting informed decision-making and confidence in the system. Based on this initial research, and the existing literature on human decision making (Smithson, 1989), it is possible to generate a number of strategies that can potentially mitigate the limitations of Type I thinking, while also supporting Type II Thinking:

- Present diverse and comprehensive information to operators, including both confirming and conflicting information, to ensure that operators are not biased towards a particular set of information. For example, in the context of medical decision-making, studies have demonstrated that providing physicians with a comprehensive set of patient information, including both confirming and conflicting information, can lead to more accurate diagnoses and treatment decisions (Eva & Norman, 2005; Lighthall & Vazquez-Guillamet, 2015). Similarly, in the field of aviation, providing pilots with a comprehensive display of flight information, including multiple sources of data and alerts, has been shown to reduce confirmation bias and improve decision-making in critical situations (Kaempf & Klein, 2017).
- Develop decision support tools that facilitate systematic evaluation of options. This has already been shown to be quite effective in a variety of contexts: From real estate appraisal (George et al., 2000), fishing management strategies (Gong et al., 2017), preferential choice problems (Todd & Benbasat, 1991) to algorithmic trading systems that provide real-time market data (Bhandari et al., 2008). Autonomous systems may encounter novel or unexpected situations where pre-programmed responses may not be effective. Type II thinking, with its deliberative and flexible nature, can be valuable in adapting to such novel situations. Integrating support systems that could “encourage” users to switch between Type I and Type II thinking modes and facilitate adaptive decision-making, can enhance the autonomy and adaptability of the system.

3.3.2 Managing cognitive workload

Cognitive workload management is a critical aspect of human decision-making and performance, particularly in high-stress situations or when individuals are required to manage multiple tasks simultaneously. Type II thinking, characterized by deliberate and effortful cognitive processing, can be overwhelmed in such situations. Designing interactions that minimize cognitive workload and cognitive switching (e.g. transitioning cognitive focus, attention, or cognitive resources from one task, activity, or mode of interaction to another) can help users effectively manage their resources. This is particularly critical when using autonomous systems in high-stress situations, such as emergency situations or critical military decision-making scenarios, where individuals may experience heightened arousal and stress, which can further impair their ability to engage in Type II thinking (Junger, 2018; Yu, 2016). There are several possible strategies to consider here:

- Reduce Perceptual and Informational Load: Perceptual load refers to the amount of visual and auditory information that users need to process (Lavie & Tsal, 1994), while informational load refers to the amount of cognitive effort required to understand and interpret information (Westbrook & Braver, 2015). Minimizing perceptual and informational load can effectively mitigate and prevent cognitive overload. Studies have demonstrated that augmenting visual displays with directional cues can significantly decrease mental workload and lead to reduced time for completing navigation tasks, and improved situation awareness (Davis, 2007). (de Melo et al., 2020), in an experiment involving an augmented reality desert survival task, showed that the use of an *embodied* assistant, allowing the use of gestures and emotions, led to higher performance by lowering cognitive burden on the decision maker.
- Develop heuristic control strategies: Studies have investigated various methods for estimating cognitive workload, including physiological indicators such as eye gaze (Aygun et al., 2022), and have explored the potential for real-time feedback to manage cognitive workload (Knisely et al., 2021; Pomranky & Wojciechowski, 2007). Autonomous systems that can “read” their users and respond to factors such as self-confidence and signs of fatigue , could not only provide better management of cognitive resources (Yuh et al., 2022), but could in fact improve hybrid team interactions and result in better performance overall (Wiltshire et al., 2014, 2022).
- Minimize Interruptions and Distractions: Interruptions and distractions can significantly impact cognitive workload and decision-making performance. Avoiding unnecessary notifications, providing options to mute or disable alerts, and designing interfaces that allow users to focus on their tasks without unnecessary interruptions can help reduce cognitive workload and prevent decision-making errors. Research has shown that simply using familiar icons, labels, and navigation patterns that align with users' mental models can go a long way in reducing cognitive switching and enhancing decision-making (Norman, 1983; Zhang & Patel, 2006).

3.3.3 Developing tailored training and skill development

Autonomous systems, such as robots, drones, and self-driving vehicles, are increasingly being utilized in complex and dynamic environments where they need to make decisions and operate autonomously. Human users of these systems often need to acquire new skills or adapt existing ones to effectively interact with and operate these autonomous systems. This process of skill development typically involves Type II thinking, which requires deliberate and effortful cognitive processing. Optimizing skill development can be achieved through various mechanisms, ranging from training programs that provide users with structured and guided learning experiences to familiarize them with autonomous systems, to simulations that create realistic and controlled environments where users can practice their skills in a safe and controlled manner, without the risk of real-world consequences. Feedback mechanisms, such as performance metrics and real-time feedback (see above), can also provide users with information about their performance and help them monitor their progress, identify areas for improvement, and refine their skills.

- In the context of unmanned vehicles, studies have shown that fostering collaboration and training among (in particular experienced) human operators through regular debriefings and team discussions can improve decision-making and situation awareness, leading to better performance and safety outcomes (Thieme & Utne, 2017). Interestingly, training and learning also seems to have an positive impact when applied to systems themselves: Through cooperative reinforcement, multiple UAVs can work together to converge on optimal control parameters faster than when working individually, using a Leader-Follower approach that coordinates their learning strategies and results in faster learning without performance loss (Jardine & Givigi, 2021).
- Providing training on cognitive biases, such as debiasing techniques and strategies, can help reduce the impact of biases on decision-making in various domains, including military decision-making, aviation, and finance (AlKhars et al., 2019; Sellier et al., 2019). Such training can empower operators to actively question their assumptions, challenge their preconceived notions, and consider alternative perspectives when interacting with autonomous systems.

Table 3. Overview of behavioral insights and their associated strategies.

Insight	Description	Strategies
Cognitive biases and Type I thinking.	Mitigate the limitations of Type I thinking, and support Type II Thinking	● Present diverse and comprehensive information to operators ● Develop decision support tools that facilitate systematic evaluation of options
Cognitive workload management	Minimize cognitive workload and switching to help users effectively manage their resources.	● Reduce Perceptual and Informational Load ● Develop heuristic control strategies ● Minimize Interruptions and Distractions
Training and skill development.	Users often need to acquire new skills or adapt existing ones to effectively interact with and operate autonomous systems.	● Foster collaboration and training among (in particular experienced) human operators ● Provide training on cognitive biases, such as debiasing techniques and strategies

3.4 Ethical considerations

Research in this area is still very much in its infancy, and it brings with it as many possibilities as it does challenges. Designers and practitioners must ensure that their methods and techniques align with ethical principles and respect the privacy, autonomy, and dignity of users. The use of behavioral science in the design of HMI should therefore always be approached with caution. Studies have found that facial recognition algorithms can be biased against certain racial or ethnic groups, leading to inaccurate identification and discrimination (Lunter, 2020). In those situations, systems could prompt users to not blindly rely on the information provided, and encourage them to engage system II thinking (see also (Nurse et al., 2022). Autonomous systems that collect or transmit personal data can potentially infringe on the privacy and autonomy of individuals. Consider the case of autonomous vehicles that collect data on passengers' movements and activities. Heuristic control strategies (see above) can be effective in mitigating operator fatigue and cognitive overload, but can also be potentially used to determine liability in accidents. We must take steps to ensure that the data collected is stored securely and used only for its intended purpose. There is also the issue of manipulation: While nudging and influencing user behavior can be beneficial in many cases, there is a fine line between gentle persuasion and unethical manipulation. We must be mindful of our responsibility to respect user autonomy and not cross into unethical practices that exploit vulnerabilities or manipulate users into making choices that are not in their best interests. Transparency, informed consent, and user empowerment should always be prioritized in the design process.

3.5 Limitations and challenges

This paper suggests a number of strategies, informed by empirical research, to help improve the design of human-machine interactions. When exploring these strategies in the context of human-machine interactions, it will be essential to assess their effectiveness while also considering any potential negative effects that might arise from their application. For instance, a key argument made in this paper is that humans mostly operate on System 1 thinking, which relies on heuristics and shortcuts. Analytical thinking (System 2) is triggered rarely due to its slower and effortful nature. A number of strategies are suggested to engage people in more analytical thinking, which is crucial to reduce overreliance on AI. However, research has already shown that there can be a trade-off: Cognitive forcing interventions that elicit or Type II thinking can be effective in reducing overreliance on AI, but tend to be perceived as less user-friendly because they require more cognitive effort (Buçinca et al., 2021). The challenge for future researchers lies in effectively evaluating and balancing the impact of strategies aimed at enhancing human-machine interactions, while being mindful of potential trade-offs and negative consequences. The engagement of System 2 thinking can indeed reduce overreliance on AI, but can also affect decision making or human-machine collaboration negatively in other ways. Addressing this challenge requires innovative approaches to find the optimal equilibrium between promoting analytical thinking and maintaining a user-friendly experience.

Another key challenge will be to develop a better understanding of the influence of individual differences, particularly with regard to cognitive biases. Such differences encompass personality traits, cognitive abilities, expertise, and decision-making styles that can significantly impact how users perceive and act upon machine learning outputs (Atchley et al., 2022; Kliegr et al., 2021). For example, certain personality traits might make individuals more susceptible to confirmation bias, while higher working memory capacity could reduce the influence of other biases. Domain experts may interpret machine learning predictions differently from non-experts due to their deeper understanding of the domain. Some research has already demonstrated how users with higher motivation to engage in effortful mental activities benefited more from the interventions forcing them to use Type II thinking than others (Buçinca et al., 2021). Addressing individual differences in the context of cognitive biases can be challenging, but is crucial for designing personalized debiasing interventions and also for exploring how education and interventions targeting individual differences can effectively debias machine learning interpretations for different user groups. Future research should focus on exploring these connections to design personalized debiasing strategies and user-centric AI systems that align with the unique characteristics of different user groups. By doing so, we can enhance the interpretability and reliability of machine learning models, fostering more informed decision-making and building user trust in AI technologies.

A great deal of behavioral economics research is focused on cognitive biases, and how type I and II thinking interact to facilitate or inhibit such biases. However, while cognitive biases are crucial aspects of human cognition affecting decision-making, other cognitive processes play a role in how users perceive and comprehend machine learning outputs. For instance, one such area of study is the examination of how users perceive and process negations in machine learning rules. The presence of negations, such as "not" or "no," in rules can introduce complexities in interpretation, potentially leading to misunderstandings or misinterpretations (Kliegr et al., 2021). Future research could focus on understanding the cognitive load associated with processing negations, the effects of user expertise and background on interpreting negated rules, and the implications of negations on user confidence in the model's predictions (see for instance (Deutsch et al., 2009; Jiang et al., 2014). By exploring the interplay between cognitive processes, expertise, and negations in rule-based machine learning models, researchers can gain insights to design more effective and user-friendly AI systems that facilitate transparent, interpretable, and comprehensible interactions between users and AI models. Such investigations can pave the way for enhanced decision-making, better trust in AI technologies, and broader applications of interpretable machine learning in various domains.

Addressing cognitive biases is a crucial aspect to ensure the reliability and fairness of decision-making processes. However, more normative work is needed to assess the seriousness of biases and prioritize their treatment, to establish standards and guidelines for what biases should be considered acceptable or problematic in XAI systems (Bertrand et al., 2022; Xiong et al., 2022). By distinguishing between "normal" biases, which are considered neutral heuristics inherent in human cognition, and "problematic" biases that distort decision-making, researchers can focus on addressing biases that have a substantial negative impact on decision outcomes. This distinction helps avoid unnecessary efforts to mitigate biases that do not significantly affect decision quality. Additionally, normative assessments can aid in prioritizing different biases based on their effects on decision accuracy, fairness, and user trust. This prioritization allows researchers to make informed decisions about which biases need to be addressed and the appropriate strategies for mitigation. Furthermore, normative work can guide researchers in making tradeoffs between different design choices in HMI. By grounding the research in normative principles, HMI can be developed to align with societal values and promote fair, trustworthy, and effective decision-making processes.

4 Conclusion

Behavioral economics can help us understand how unpredictable, and often irrational, the behavior of human operators can be. In particular the tension between type I (fast, automatic but potentially biased) and type II (controlled, conscious, rational but limited in capacity) thinking could be relevant for design practitioners. This paper provided a number of ideas and possible strategies to think about when we design and shape how people and autonomous systems work together. By applying these strategies, we can potentially make HMI related interfaces, decision aids, and training programs that lead to better performance, more trust, and fewer mistakes and accidents. Given that research in this area is often still in early stages, it is, however, important to continue to investigate and explore how such insights can be integrated into our design practices. At the same time, it is imperative that we approach this process with critical thinking, ethical considerations, and a deep understanding of the limitations and nuances of the research findings. By doing this, we can get the most out of autonomous systems, keep the risks and limits of human-machine interaction to a minimum and create designs that not only delight and engage users, but also respect their autonomy, diversity, and well-being.

References

AlKhars, M., Evangelopoulos, N., Pavur, R., & Kulkarni, S. (2019). Cognitive biases resulting from the representativeness heuristic in operations management: an experimental investigation. *Psychology Research and Behavior Management*, *12*, 263–276. https://doi.org/10.2147/PRBM.S193092

Atchley, A., Barr, H. M., O'Hear, E., Weger, K., Mesmer, B., Gholston, S., & Tenhundfeld, N. (2022). *Trust in Systems: Identification of 17 Unresolved Research Questions and the Highlighting of Inconsistencies*. https://doi.org/10.31219/osf.io/r4gpd

Aygun, A., Nguyen, T., Haga, Z., Aeron, S., & Scheutz, M. (2022). Investigating Methods for Cognitive Workload Estimation for Assistive Robots. *Sensors* , *22*(18). https://doi.org/10.3390/s22186834

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). *How cognitive biases affect XAI-Assisted decision-making: A systematic review* (AAAI; ACM SIGAI, trans.). 5th AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, AIES 2022. Association for Computing Machinery, Inc. https://doi.org/10.1145/3514094.3534164

Bewley, W. L., Lee, J. J., Jones, B., & Cai, H. (2014). Assessing cognitive readiness in a simulation-based training environment. In *Teaching and Measuring Cognitive Readiness* (Vol. 9781461475798, pp. 253–278). Springer US. https://doi.org/10.1007/978-1-4614-7579-8_14

Bhandari, G., Hassanein, K., & Deaves, R. (2008). Debiasing investors with decision support systems: An experimental investigation. *Decision Support Systems*, *46*(1), 399–410. https://doi.org/10.1016/j.dss.2008.07.010

Broadbent, D. (1990). A Problem Looking for Solutions. *Psychological Science*, *1*(4), 235–239. https://doi.org/10.1111/j.1467-9280.1990.tb00206.x

Brown, T. (2008). Design thinking. *Harvard Business Review*, *86*(6), 84–92, 141. https://www.ncbi.nlm.nih.gov/pubmed/18605031

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1). https://doi.org/10.1145/3449287

Cash, P., Gamundi, X. V., Echstrøm, I., & Daalhuizen, J. (2022). Method Use in Behavioural Design: What, How, and Why? *International Journal of Design*, *16*(1). https://doi.org/10.57698/v16i1.01

Choi, J. K., & Ji, Y. G. (2015). Investigating the Importance of Trust on Adopting an Autonomous Vehicle. *International Journal of Human–Computer Interaction*, *31*(10), 692–702. https://doi.org/10.1080/10447318.2015.1070549

Cialdini, R. B., & Jacobson, R. P. (2021). Influences of social norms on climate change-related behaviors. *Current Opinion in Behavioral Sciences*, *42*, 1–8. https://doi.org/10.1016/j.cobeha.2021.01.005

Cross, E. S., & Ramsey, R. (2021). Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions. *Trends in Cognitive Sciences*, *25*(3), 200–212. https://doi.org/10.1016/j.tics.2020.11.009

Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2020). Wearable Reasoner: Towards Enhanced Human Rationality Through A Wearable Device With An Explainable AI Assistant. *Proceedings of the Augmented Humans International Conference*, Article Article 23. https://doi.org/10.1145/3384657.3384799

de Melo, C. M., Kim, K., Norouzi, N., Bruder, G., & Welch, G. (2020). Reducing Cognitive Load and Improving Warfighter Problem Solving With Intelligent Virtual Assistants. *Frontiers in Psychology*, *11*, 554706. https://doi.org/10.3389/fpsyg.2020.554706

De Regt, A., & Gagnon, E. (2020). Rethinking how humans and machines make sense together. *Proc. 26th Amer. Conf. Inf. Syst*. https://scholar.archive.org/work/kyzgp42jjjdztafqqkxhslinhi/access/wayback/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1422&context=amcis2020

Deterding, S. (2012). Gamification: designing for motivation. *Interactions*, *19*(4), 14–17. https://doi.org/10.1145/2212877.2212883

Deutsch, R., Kordts-Freudinger, R., Gawronski, B., & Strack, F. (2009). Fast and fragile: A new look at the automaticity of negation processing. *Experimental Psychology*, *56*(6), 434–446. https://doi.org/10.1027/1618-3169.56.6.434

Effie Lai-Chong Law, Ebba Thora Hvannberg, Gilbert Cockton (Ed.). (n.d.). *Maturing Usability*. Springer London. https://doi.org/10.1007/978-1-84628-941-5

Endsley, M. R. (1995). Measurement of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 65–84. https://doi.org/10.1518/001872095779049499

Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors*, *37*(2), 381–394. https://doi.org/10.1518/001872095779064555

Eva, K. W., & Norman, G. R. (2005). Heuristics and biases--a biased perspective on clinical reasoning [Review of *Heuristics and biases--a biased perspective on clinical reasoning*]. *Medical Education*, *39*(9), 870–872. https://doi.org/10.1111/j.1365-2929.2005.02258.x

Ferreira, J. M., Acuña, S. T., Dieste, O., Vegas, S., Santos, A., Rodríguez, F., & Juristo, N. (2020). Impact of usability mechanisms: An experiment on efficiency, effectiveness and user satisfaction. *Information and Software Technology*, *117*(106195), 106195. https://doi.org/10.1016/j.infsof.2019.106195

George, J. F., Duffy, K., & Ahuja, M. (2000). Countering the anchoring and adjustment bias with decision support systems. *Decision Support Systems*, *29*(2), 195–206. https://doi.org/10.1016/s0167-9236(00)00074-9

Gong, M., Lempert, R., Parker, A., Mayer, L. A., Fischbach, J., Sisco, M., Mao, Z., Krantz, D. H., & Kunreuther, H. (2017). Testing the scenario hypothesis: An experimental comparison of scenarios and forecasts for decision support in a complex decision environment. *Environmental Modelling and Software[R]*, *91*, 135–155. https://doi.org/10.1016/j.envsoft.2017.02.002

Hopko, S., Wang, J., & Mehta, R. (2022). Human Factors Considerations and Metrics in Shared Space Human-Robot Collaboration: A Systematic Review. *Frontiers in Robotics and AI*, *9*, 799522. https://doi.org/10.3389/frobt.2022.799522

Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019). History and future of human-automation interaction. *International Journal of Human-Computer Studies*, *131*, 99–107. https://doi.org/10.1016/j.ijhcs.2019.05.006

Jardine, P. T., & Givigi, S. (2021). Improving Control Performance of Unmanned Aerial Vehicles through Shared Experience. *Journal of Intelligent and Robotic Systems*, *102*(3), 68. https://doi.org/10.1007/s10846-021-01387-1

Jeffries, R., & Wixon, D. (2007). *Maturing Usability: Quality in Software, Interaction and Value (Human–Computer Interaction Series)* (E. L.-C. Law, E. Hvannberg, & G. Cockton (Eds.); 2008th ed.). Springer.

Jiang, Z.-Q., Li, W.-H., Liu, Y., Luo, Y.-J., Luu, P., & Tucker, D. M. (2014). When affective word valence meets linguistic polarity: Behavioral and ERP evidence. *Journal of Neurolinguistics*, *28*, 19–30. https://doi.org/10.1016/j.jneuroling.2013.11.001

Junger, P. M. (2018). *The effects of hypervigilance on decision-making during critical incidents*. Naval Postgraduate School. https://apps.dtic.mil/sti/citations/AD1065390

Kaempf, G. L., & Klein, G. (2017). Aeronautical Decision Making: The next generation. In *Aviation Psychology in Practice* (pp. 223–254). Routledge. https://doi.org/10.4324/9781351218825-11

Kahneman, D. (2013). *Thinking, Fast and Slow* (1st ed.). Farrar, Straus and Giroux. https://www.amazon.com/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374533555

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*, 103458. https://doi.org/10.1016/j.artint.2021.103458

Knisely, B. M., Joyner, J. S., & Vaughn-Cooke, M. (2021). Cognitive task analysis and workload classification. *MethodsX*, *8*, 101235. https://doi.org/10.1016/j.mex.2021.101235

Krausman, A., Neubauer, C., Forster, D., Lakhmani, S., Baker, A. L., Fitzhugh, S. M., Gremillion, G., Wright, J. L., Metcalfe, J. S., & Schaefer, K. E. (2022). Trust Measurement in Human-Autonomy Teams: Development of a Conceptual Toolkit. *ACM Transactions on Human-Robot Interaction*, *11*(3), 1–58. https://doi.org/10.1145/3530874

Kuhail, M. A., Abu Shawar, B., & Hammad, R. (2023). *Trends, Applications, and Challenges of Chatbot Technology*. IGI Global. https://play.google.com/store/books/details?id=Rv2uEAAAQBAJ

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, *56*(2), 183–197. https://doi.org/10.3758/bf03213897

Lee, J.-G., Kim, K. J., Lee, S., & Shin, D.-H. (2015). Can Autonomous Vehicles Be Safe and Trustworthy? Effects of Appearance and Autonomy of Unmanned Driving Systems. *International Journal of Human–Computer Interaction*, *31*(10), 682–691. https://doi.org/10.1080/10447318.2015.1070547

Lighthall, G. K., & Vazquez-Guillamet, C. (2015). Understanding Decision Making in Critical Care. *Clinical Medicine & Research*, *13*(3-4), 156–168. https://doi.org/10.3121/cmr.2015.1289

Lunter, J. (2020). Beating the bias in facial recognition technology. *Biometric Technology Today*, *2020*(9), 5. https://doi.org/10.1016/S0969-4765(20)30122-3

Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human-Autonomy Teaming: Definitions, Debates, and Directions. *Frontiers in Psychology*, *12*, 589585. https://doi.org/10.3389/fpsyg.2021.589585

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *The Journal of Applied Psychology*, *85*(2), 273–283. https://doi.org/10.1037/0021-9010.85.2.273

Mayer, R. E. (2014). What Problem Solvers Know: Cognitive Readiness for Adaptive Problem Solving. In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Teaching and Measuring Cognitive Readiness* (pp. 149–160). Springer US. https://doi.org/10.1007/978-1-4614-7579-8_8

Meppelink, C. S., Smit, E. G., Fransen, M. L., & Diviani, N. (2019). "I was Right about Vaccination": Confirmation Bias and Health Literacy in Online Health Information Seeking. *Journal of Health Communication*, *24*(2), 129–140. https://doi.org/10.1080/10810730.2019.1583701

Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(1). https://doi.org/10.1073/pnas.2107346118

Norman, D. (2013). *The Design of Everyday Things: Revised and Expanded Edition*. Hachette UK.

Norris, J. N. (2018). Human Factors in Military Maritime and Expeditionary Settings: Opportunity for Autonomous Systems? *Advances in Human Factors in Robots and Unmanned Systems*, 139–147. https://doi.org/10.1007/978-3-319-60384-1_14

Nurse, M. S., Ross, R. M., Isler, O., & Van Rooy, D. (2022). Analytic thinking predicts accuracy ratings and willingness to share COVID-19 misinformation in Australia. *Memory & Cognition*, *50*(2), 425–434. https://doi.org/10.3758/s13421-021-01219-5

Pomranky, R. A., & Wojciechowski, J. Q. (2007). *Determination of mental workload during operation of multiple unmanned systems*. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING …. https://apps.dtic.mil/sti/citations/ADA474506

Schelble, B. G., Flathmann, C., & McNeese, N. (2020). Towards Meaningfully Integrating Human-Autonomy Teaming in Applied Settings. *Proceedings of the 8th International Conference on Human-Agent Interaction*, 149–156. https://doi.org/10.1145/3406499.3415077

Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing Training Improves Decision Making in the Field. *Psychological Science*, *30*(9), 1371–1379. https://doi.org/10.1177/0956797619861429

Smithson, M. (1989). Ignorance and uncertainty: Emerging paradigms. *Cognitive Science*, *393*. https://doi.org/10.1007/978-1-4612-3628-3

Sparrow, R. (2009). Building a better warbot: ethical issues in the design of unmanned systems for military applications. *Science and Engineering Ethics*, *15*(2), 169–187. https://doi.org/10.1007/s11948-008-9107-0

Springer, A., & Whittaker, S. (2019). Progressive disclosure: empirically motivated approaches to designing effective transparency. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 107–120. https://doi.org/10.1145/3301275.3302322

Stea, S., & Pickering, G. J. (2019). Optimizing Messaging to Reduce Red Meat Consumption. *Environmental Communication*, *13*(5), 633–648. https://doi.org/10.1080/17524032.2017.1412994

Strengers, Y. (2013). *Smart Energy Technologies in Everyday Life: Smart Utopia?* Springer. https://market.android.com/details?id=book-HU4hAQAAQBAJ

Thieme, C. A., & Utne, I. B. (2017). A risk model for autonomous marine systems and operation focusing on human–autonomy collaboration. *Proceedings of the Institution of Mechanical Engineers. Part O, Journal of Risk and Reliability*, *231*(4), 446–464. https://doi.org/10.1177/1748006x17709377

Todd, P., & Benbasat, I. (1991). An Experimental Investigation of the Impact of Computer Based Decision Aids on Decision Making Strategies. *Information Systems Research*, *2*(2), 87–115. https://doi.org/10.1287/isre.2.2.87

Traut, J. (2023). Morgan Housel: The psychology of money: timeless lessons on wealth, greed, and happiness (Harriman House, 2020). *Financial Markets and Portfolio Management*. https://doi.org/10.1007/s11408-022-00424-9

Voyer, B. G. (2015). "Nudging" behaviours in healthcare: insights from behavioural economics. *British Journal of Healthcare Management*, *21*(3), 130–135. https://doi.org/10.12968/bjhc.2015.21.3.130

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Article Paper 601. https://doi.org/10.1145/3290605.3300831

Wenker, K. (2022). A Systematic Literature Review on Persuasive Technology at the Workplace. In *arXiv [cs.HC]*. arXiv. http://arxiv.org/abs/2201.00329

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective & Behavioral Neuroscience*, *15*(2), 395–415. https://doi.org/10.3758/s13415-015-0334-y

Wiltshire, T. J., Rosch, K., Fiorella, L., & Fiore, S. M. (2014). Training for Collaborative Problem Solving: Improving Team Process and Performance through Metacognitive Prompting. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting Human Factors and Ergonomics Society. Meeting*, *58*(1), 1154–1158. https://doi.org/10.1177/1541931214581241

Wiltshire, T. J., van Eijndhoven, K., Halgas, E., & Gevers, J. M. P. (2022). Prospects for Augmenting Team Interactions with Real-Time Coordination-Based Measures in Human-Autonomy Teams. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12606

Xiong, W., Fan, H., Ma, L., & Wang, C. (2022). Challenges of human—machine collaboration in risky decision-making. *Frontiers of Engineering Management*, *9*(1), 89–103. https://doi.org/10.1007/s42524-021-0182-0

Yuh, M. S., Byeon, S., Hwang, I., & Jain, N. (2022). A Heuristic Strategy for Cognitive State-based Feedback Control to Accelerate Human Learning. *IFAC-PapersOnLine*, *55*(41), 107–112. https://doi.org/10.1016/j.ifacol.2023.01.111

Yu, R. (2016). Stress potentiates decision biases: A stress induced deliberation-to-intuition (SIDI) model. *Neurobiology of Stress*, *3*, 83–95. https://doi.org/10.1016/j.ynstr.2015.12.006

Zaharias, P., & Poulymenakou, A. (2006). Implementing learner-centred design: The interplay between usability and instructional design practices. *Interactive Technology and Smart Education*, *3*(2), 87–100. https://doi.org/10.1108/17415650680000055

About the Authors:

Prof. Dirk Van Rooy: academic and consultant active in Europe and Australia, currently based at the Antwerp Center of Responsible AI (ACRAI). His research is focused on human-AI interaction, psychology and decision-making in industry and defence applications.