frontiers | Research Topics

Distributed cognition in learning and behavioral change – based on human and artificial intelligence

Edited by
Dietrich Albert, Tomoko Kojiri, Xiangen Hu and Paul Seitlinger

Published in
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-4231-6
DOI 10.3389/978-2-8325-4231-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Distributed cognition in learning and behavioral change – based on human and artificial intelligence

Topic editors

Dietrich Albert — University of Graz, Austria
Tomoko Kojiri — Kansai University, Japan
Xiangen Hu — University of Memphis, United States
Paul Seitlinger — University of Vienna, Austria

Citation

Albert, D., Kojiri, T., Hu, X., Seitlinger, P., eds. (2024). *Distributed cognition in learning and behavioral change – based on human and artificial intelligence*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4231-6

Table of contents

frontiers | Frontiers in Artificial Intelligence

PERSPECTIVE
published: 14 July 2022
doi: 10.3389/frai.2022.908261

Supporting Cognition With Modern Technology: Distributed Cognition Today and in an AI-Enhanced Future

Sandra Grinschgl and Aljoscha C. Neubauer*

Institute of Psychology, University of Graz, Graz, Austria

In the present article, we explore prospects for using artificial intelligence (AI) to distribute cognition *via* cognitive offloading (i.e., to delegate thinking tasks to AI-technologies). Modern technologies for cognitive support are rapidly developing and increasingly popular. Today, many individuals heavily rely on their smartphones or other technical gadgets to support their daily life but also their learning and work. For instance, smartphones are used to track and analyze changes in the environment, and to store and continually update relevant information. Thus, individuals can offload (i.e., externalize) information to their smartphones and refresh their knowledge by accessing it. This implies that using modern technologies such as AI empowers users *via* offloading and enables them to function as always-updated knowledge professionals, so that they can deploy their insights strategically instead of relying on outdated and memorized facts. This AI-supported offloading of cognitive processes also saves individuals' internal cognitive resources by distributing the task demands into their environment. In this article, we provide (1) an overview of empirical findings on cognitive offloading and (2) an outlook on how individuals' offloading behavior might change in an AI-enhanced future. More specifically, we first discuss determinants of offloading such as the design of technical tools and links to metacognition. Furthermore, we discuss benefits and risks of cognitive offloading. While offloading improves immediate task performance, it might also be a threat for users' cognitive abilities. Following this, we provide a perspective on whether individuals will make heavier use of AI-technologies for offloading in the future and how this might affect their cognition. On one hand, individuals might heavily rely on easily accessible AI-technologies which in return might diminish their internal cognition/learning. On the other hand, individuals might aim at enhancing their cognition so that they can keep up with AI-technologies and will not be replaced by them. Finally, we present own data and findings from the literature on the assumption that individuals' personality is a predictor of trust in AI. Trust in modern AI-technologies might be a strong determinant for wider appropriation and dependence on these technologies to distribute cognition and should thus be considered in an AI-enhanced future.

Keywords: technology, artificial intelligence (AI), distributed cognition, cognitive offloading, trust

OPEN ACCESS

Edited by:
Xiangen Hu,
University of Memphis, United States

Reviewed by:
Paul Lefrere,
Clear Communication Associates,
United Kingdom

****Correspondence:***
Sandra Grinschgl
sandra.grinschgl@uni-graz.at
orcid.org/0000-0001-6666-9426

Specialty section:
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

Received: *30 March 2022*
Accepted: *24 June 2022*
Published: *14 July 2022*

Citation:
Grinschgl S and Neubauer AC (2022) Supporting Cognition With Modern Technology: Distributed Cognition Today and in an AI-Enhanced Future. Front. Artif. Intell. 5:908261. doi: 10.3389/frai.2022.908261

INTRODUCTION

Today, modern technologies are an indispensable part of peoples' lives and it is hard to image living without smartphones and other technical gadgets. Currently, around 1.3 billion smartphones are sold worldwide per year (idc.com, 2022) and, for instance, in Austria about 83% of individuals above 15 years own a smartphone (KMU Forschung Austria, 2020). People are using their technical devices for several reasons, such as staying in contact, using the internet, taking pictures, or playing games. Furthermore, technical devices can be used to externalize cognitive processes, which is referred to as *cognitive offloading* (Risko and Gilbert, 2016). Individuals offload cognitive processes by, for instance, relying on navigation applications instead relying on one's own spatial abilities or by storing appointments or shopping lists in their smartphones instead of memorizing them. Also, modern technologies can be used to access up-to-date knowledge and thus individuals do not need to rely on outdated, internally memorized information. Indeed, the majority of adults indicates using technical devices regularly or even very often as external memory stores (Finley et al., 2018). Also, empirical studies show that individuals flexibly distribute cognitive demands between internal and external resources when solving problems (e.g., Cary and Carlson, 2001). Distributed cognition using modern technologies should thereby facilitate task performance.

Besides the mentioned basic applications of a technical device, also smart applications that are based on artificial intelligence (AI) might be used for offloading. For instance, smart speakers could be used to store appointments and to be reminded of them. Due to the effortless interaction with such smart applications distributed cognition might reach a new all-time high in the coming years. Here, we aim at discussing this development of distributed cognition with modern technologies. First, we give an overview of recent findings regarding cognitive offloading. We discuss determinants that foster the offloading of cognitive processes in modern technologies as well as possible benefits and risks of offloading. Second, we provide an outlook on how distributed cognition (and thus cognitive offloading) might change in the future–when AI-technologies are used on a daily basis by many people–especially with regard to the educational sector. Furthermore, we discuss whether distributed cognition will actually increase in the future or whether people will rather aim at enhancing their internal cognitive abilities. One crucial factor for using AI-technologies to distribute cognition might be peoples' trust in these technologies. Therefore, we also describe personality traits as potential predictors of trust in AI by summarizing findings from the literature and presenting our own data. Finally, we argue for more psychological research that could support and inform the development of modern technologies.

OVERVIEW OF COGNITIVE OFFLOADING WITH MODERN TECHNOLOGIES

While cognitive offloading comes in many different forms (e.g., for short term memory offloading see Meyerhoff et al., 2021; for navigation offloading see Fenech et al., 2010), investigations on the offloading of cognitive processes into modern technology can be summarized into two lines of research: (1) the determinants of cognitive offloading and (2) the consequences of offloading behavior.

Modern technologies could be used for offloading whenever they are available, but offloading might be applied more or less depending on certain conditions. On the one hand, offloading was shown to depend on external factors such as the design of technical tools (e.g., Gray et al., 2006; Grinschgl et al., 2020) or characteristics of the to-be-processed information (e.g., Schönpflug, 1986; Hu et al., 2019). Regarding tool design, studies observed that individuals offload more cognitive processes when the offloading process (i.e. the interaction with a technical tool) is fast vs. associated with temporal delays (e.g., Gray et al., 2006; Waldron et al., 2011; Grinschgl et al., 2020). Similarly, also when interacting with technical tools requires less vs. more operational steps offloading increases (e.g., O'Hara and Payne, 1998; Cary and Carlson, 2001). Furthermore, Grinschgl et al. (2020) observed that when participants performed a task on a tablet using its touch function, they offloaded more working memory processes than when using a computer mouse. Regarding the characteristics of the to-be-processed information, it was shown that offloading increases when a task and accompanying information is more complex, relevant, or difficult (e.g., Schönpflug, 1986; Hu et al., 2019). Additionally, a larger amount of to-be-processed information fosters offloading behavior (e.g., Gilbert, 2015a; Arreola et al., 2019). Overall, these external factors suggest that the distribution of cognitive processes on internal and external resources depends on situational cost-benefit considerations (e.g., Gray et al., 2006; Grinschgl et al., 2020).

On the other hand, internal factors such as individuals' cognitive abilities and metacognitive beliefs can impact offloading. Studies observed more offloading when one's working memory capacity is lower (e.g., Gilbert, 2015b; Meyerhoff et al., 2021). Moreover, individuals commonly offload more when they believe that their internal performance is worse (Gilbert, 2015b; Boldt and Gilbert, 2019; but see Grinschgl et al., 2021a for conflicting results). To our best knowledge, an investigation of other individual differences with regard to cognitive offloading is still lacking (e.g., there is no research on the interplay between personality and offloading).

Together, these determinants of offloading behavior suggest that individuals do not maximally offload under all circumstances, but instead especially the easy and fast access to modern technology fosters offloading (e.g., Grinschgl et al., 2020). With an increase in offloading due to modern technologies, the question arises whether offloading is accompanied by positive and/or negative consequences. Cognitive offloading was shown to improve immediate task performance by accelerating it and/or reducing errors (e.g., Boldt and Gilbert, 2019; Grinschgl et al., 2021b). Furthermore, studies showed that offloading improves simultaneous secondary task performance (Grinschgl et al., 2022) and later performance of unrelated tasks (Storm and Stone, 2015; Runge et al., 2019). Hence, it is assumed that cognitive offloading releases

internal cognitive resources that can be devoted to other, simultaneous or subsequent tasks. Furthermore, the retrieval and storage of information using modern technologies might help individuals to refresh their internal knowledge and thus to act as always-updated knowledge professionals.

Importantly, cognitive offloading is also accompanied by risks. In three experiments, Grinschgl et al. (2021b) observed a trade-off for cognitive offloading: while the offloading of working memory processes increased immediate task performance, it also decreased subsequent memory performance for the offloaded information. Similarly, the offloading of spatial processes by using a navigation device impairs spatial memory (i.e., route learning and subsequent scene recognition; Fenech et al., 2010). Thus, information stored in a technical device might be quickly forgotten (for an intentional/directed forgetting account see Sparrow et al., 2011; Eskritt and Ma, 2014) or might not be processed deeply enough so that no long-term memory representations are formed (cf. depth of processing theories; Craik and Lockhart, 1972; Craik, 2002). In addition to detrimental effects of offloading on (long-term) memory, offloading hinders skill acquisition (Van Nimwegen and Van Oostendorp, 2009; Moritz et al., 2020) and harms metacognition (e.g., Fisher et al., 2015, 2021; Dunn et al., 2021); e.g. the use of technical aids can inflate one's knowledge. In Dunn et al. (2021), the participants had to answer general knowledge questions by either relying on their internal resources or additionally using the internet. Metacognitive judgments showed that participants were overconfident when they are allowed to use the internet. Similarly, Fisher et al. (2021) concluded that searching for information online leads to a misattribution of external information to internal memory.

To summarize, cognitive offloading is accompanied by both–benefits and risks. While it can improve immediate task performance, it might also be accompanied by detrimental long-term effects. However, the investigated time-frames were rather short (effects over hours or days). Thus, it remains unclear how offloading might impact cognition over the lifespan (for a discussion see Cecutti et al., 2021). While these authors see humans' development with modern technologies rather positive, others pose modern technologies as a threat for humans (e.g., Carr, 2008; Spitzer, 2012). If distributing cognition actually is a blessing or a threat for human cognition cannot be answered here, but we will provide a brief outlook on how distributed cognition might be affected by the rise of AI.

OUTLOOK ON DISTRIBUTED COGNITION IN AN AI-ENHANCED FUTURE

As the use of many AI-technologies such as smart speakers appears quite effortless, they might be used to easily store appointments, take notes, retrieve up-to-date knowledge, or to perform other cognitive tasks (e.g., calculating, navigating). AI-technologies might thus be the "future" of distributed cognition by replacing classical offloading tools. With this prospect, it is important to consider how the omnipresence of AI as offloading tools might impact human cognition and how we–as humans–might foster a worthwhile integration of AI into our life.

To date, cognitive offloading research has shown positive and negative consequences of using modern technology to distribute demands on internal and external resources. However, these studies did not target the use of AI-technologies as offloading tools, but standard technologies such as tablets/computers without AI applications. To our best knowledge, research investigating distributed cognition in the context of AI is lacking. While we see a high potential for new studies on this matter, we also think that previous results can be transferred to the current and future use of AI. Therefore, AI-technologies should be used with caution as they might diminish cognitive abilities such as learning and memory–consequences that are especially relevant when it comes to children's education.

Studies suggest that even young children offload cognitive processes instead of completely relying on their internal resources (e.g., Armitage et al., 2020; Bulley et al., 2020). Thus, already in crucial learning phases during childhood tools are used to distribute cognitive demands onto internal and external resources. Such offloading behavior might increase with the ever earlier access to modern AI-technologies in smartphones and computers. Especially regarding education, it must be discussed whether there should be a "ban" of cognitive offloading due to potential detrimental effects thereof or whether students need to learn how to properly use technical tools without causing harm for their cognition (cf. Bearman and Luckin, 2020; Dawson, 2020). In line with these authors, we advocate to teach students how to use technical devices so that they satisfy their needs but to not (unintentionally) harm cognition. For instance, students need to learn how to differentiate between their own knowledge and externally stored knowledge, so that the effect of inflated knowledge is avoided. Furthermore, students should be made aware of their offloading behavior and that they won't be able to access their technical tools in critical situations such as during exams. This is especially important as a study showed that cognitive offloading was not detrimental for long-term memory when participants were forced to offload but also were instructed to internally memorize the relevant information (Grinschgl et al., 2021b). Thus, offloading is not always detrimental for building long-term memory representations and instead detriments of offloading might be compensated by proper learning instructions. Additionally, modern (AI-)technologies can benefit education by providing students with automated feedback to improve learning (Bearman and Luckin, 2020). Hence, the availability of modern (AI)-technologies is accompanied by both benefits and risks.

One alternative to strongly relying on AI to perform demanding tasks, might be the enhancement of one's own cognition. Especially the Transhumanism movement in philosophy proposes the enhancement of human cognition, such as intelligence, so that we are able to solve global problems (e.g., the climate crisis; Liao et al., 2012; Sorgner, 2020). This enhancement should be achieved by enhancement methods such as taking smart drugs, stimulating the brain, or modifying genes (Bostrom and Sandberg, 2009). However, so far the effects of most enhancement methods are at best moderate (Hills and

Hertwig, 2011; Jaušovec and Pahor, 2017). The future might bring new possibilities to foster human enhancement and this might enable humans to compete with AI (for a discussion on the implications of human enhancement and AI see Neubauer, 2021).

As human enhancement is not a promising strategy to become smarter yet, individuals might rather rely on the available technologies to improve their performance. As outlined before, individuals do not rely on technical tools all the time, but the distribution of cognitive processes onto internal and external resources rather depends on factors such as tool design, one's abilities, or metacognitive beliefs. Another factor that might strongly influence the reliance on AI might be the trust in these technologies. In a recent review, Matthews et al. (2021) identified trust as a major factor when it comes to human-machine interaction and suggested that trust in AI should be systematically investigated as humans are approaching a technology-enhanced future. Trust can be seen as the specific beliefs about technology and the willingness to rely on technology in risky situations (Siau and Wang, 2018). Thus, trust might determine if and how individuals interact with AI (see also Glikson and Woolley, 2020; Chong et al., 2022) for distributed cognition. The question arises whether there are individual differences when it comes to trusting AI. An important source for individual differences might be individuals' personality such as the Big 5 traits (Hoff and Bashir, 2015; Matthews et al., 2021). Therefore, we briefly summarize research investigating personality traits as potential predictors of trust in AI and present our own data on this matter.

On one hand, several studies investigated trust regarding specific AI-technologies. Li et al. (2020) consistently observed correlations between openness and trust in automated driving. While the participants showed lower trust in automated driving with higher openness in a questionnaire, they also showed a higher monitoring frequency, more frequent, and earlier and longer "take overs" in an automated driving simulator. Individuals high in openness might strive for more intellectually demanding tasks and thus do not heavily rely on automated systems (but see Zhang et al., 2020, for conflicting results). Additionally, higher extraversion was related to less trust in the questionnaire but no other effects were observed (Li et al., 2020). In contrast, Kraus et al. (2021) did not observe a relationship between openness and trust in automated driving, but indirect effects of neuroticism, extraversion, and agreeableness. In a path model, higher neuroticism was related to less affinity for technology. A higher affinity for technology was positively related to trust in automated driving (see also Zhang et al., 2020). Moreover, higher extraversion and agreeableness were related to more interpersonal trust which was related to more trust in automated driving. These findings suggest that there is a common factor underlying both trust in humans and trust in automated systems (Kraus et al., 2021). Besides automated driving, Sharan and Romano (2020) investigated trust in AI by providing participants with suggestions from an AI-algorithm when making decisions in a card game and found none of the Big 5 traits correlated to any trust indicators.

Other studies assessed general trust in AI- (or related) technologies. For instance, Chien et al. (2016) observed that higher agreeableness and conscientiousness were related to more trust in automated systems, with no significant relationships for the other Big 5 traits. Merrit and Ilgen (2008) observed that extraversion is positively related to the propensity to trust machines; additionally, age was negatively related to trust in machines. In our own, exploratory study ($N = 467$; for details see **Supplementary Material**), we assessed general trust in AI by a 7-item questionnaire. Additionally, we measured participants sociodemographic data (age, gender, and education), Big 5 traits and facets, and their interpersonal trust. The correlations of these variables with general trust in AI can be found in **Table 1**. In contrast to single previous studies (e.g., Chien et al., 2016), we did not observe any relationships between trust in AI and Big 5 traits as well as their facets. Additionally, interpersonal trust was not related to trust in AI. We, thus, cannot confirm the findings of Kraus et al. (2021), suggesting trust in others and in technology are positively correlated. However, in line with Merrit and Ilgen (2008), we observed a negative relationship between age and trust in AI as well as no gender effects. Older adults seem to have less trust in AI, potentially due to less experience with these systems and modern technology in general (but see Zhang et al., 2020, for diverging results).

To summarize, results regarding personality traits as predictors of trust in AI seem rather inconsistent and depend on the targeted AI-technology (see also Schäfer et al., 2016). We thus urge for more systematic research to identify personality traits and other factors that might predict the trust in and

TABLE 1 | Correlations between sociodemographic data, Big 5 traits and facets, interpersonal trust, and general trust in AI.

	Trust in AI
Age	−0.17*
Gender	0.04
Education	0.09
Openness	−0.02
Openness–aesthetics	−0.01
Openness–ideas	−0.02
Agreeableness	0.02
Agreeableness–altruism	0.03
Agreeableness–concession	0.03
Conscientiousness	−0.12
Conscientiousness order	−0.12
Conscientiousness–self-discipline	−0.11
Extraversion	0.06
Extraversion–assertiveness	0.06
Extraversion–activity	0.04
Neuroticism	−0.07
Neuroticism–anxiety	−0.05
Neuroticism–depression	−0.08
Interpersonal trust	0.08

*N = 467; for gender N = 466 because of the exclusion of one non-binary participant; women are coded as 0, men are coded as 1; significance levels are Bonferroni-Holm corrected; *p < 0.05.*

use of AI-technologies in general but also for distributed cognition more specifically. Besides personality, other factors might play a major role for trusting and using AI. For instance, the perceived usability of AI-technologies, computer anxiety, task characteristics, transparency, or perceived intelligence of AI might determine if and how AI-technologies are used (for an overview and further differentiation see Glikson and Woolley, 2020; Kaplan et al., 2021; Matthews et al., 2021).

DISCUSSION

While smartphones and other technical gadgets are already frequently used to distribute cognition, the rise of easy-to-use AI-technologies might further foster the offloading of cognitive processes. This development urges a need to investigate consequences of using AI for distributed cognition and to inform the public about these consequences. Although AI is already often discussed by information scientists, politics, and the general public, psychologists and psychological findings are barely integrated into these discussions. We see a high potential for psychological research to inform the public about potentials of modern technologies but also about accompanied risks. Moreover, identifying individuals that would easily rely on AI (e.g., due to a high trust in these systems) could help in specifically targeting these individuals when it comes to potential negative consequences of heavily relying on technology. Thus, we argue for more individual differences research to systematically investigate which factors (e.g., personality) predict trust in and use of AI. Growing up with modern technologies will likely affect our cognition as well as our attitude toward technologies. Such findings should be considered both when designing modern technologies and when using them in different situations (e.g., in private life, educational settings)–already now and in an AI-enhanced future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors upon request, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the University Graz. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SG: conceptualization, methodology, data curation, data analysis, and writing—original draft. AN: conceptualization, methodology, and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

The authors acknowledge the financial support by the University of Graz for the Open Access Publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.908261/full#supplementary-material

REFERENCES

Armitage, K. L., Bulley, A., and Redshaw, J. (2020). Developmental origins of cognitive offloading. *Proc. R. Soc. B* 287:20192927. doi: 10.1098/rspb.2019.2927

Arreola, H., Flores, A. N., Latham, A., MacNew, H., and Vu, K.-P. L. (2019). Does the use of tablets lead to more information being recorded and better recall in short-term memory tasks? In: *Proceedings of the 21st HCI International Conference, Human Interface and Management of Information* Cham: Springer. doi: 10.1007/978-3-030-22660-2_20

Bearman, M., and Luckin, R. (2020). "Preparing university assessment for a world with AI: Tasks for human intelligence," In: *Re-Imagining University Assessment in a Digital World,* eds M. Bearman, P. Dawson, R. Ajjawi, J. Tai, and D. Boud (Cham: Springer). doi: 10.1007/978-3-030-41956-1_5

Boldt, A., and Gilbert, S. (2019). Confidence guides spontaneous cognitive offloading. *Cognit. Res.* 4:45. doi: 10.1186/s41235-019-0195-y

Bostrom, N., and Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Sci. Eng. Ethics* 15, 311–341. doi: 10.1007/s11948-009-9142-5

Bulley, A., McCarthy, T., Gilbert, S. J., Suddendorf, T., and Redshaw, J. (2020). Children devise and selectively use tools to offload cognition. *Curr. Biol.* 17, 3457–3464.e3. doi: 10.1016/j.cub.2020.06.035

Carr, N. (2008). *Is Google Making Us Stupid?* The Atlantic. https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/

Cary, M., and Carlson, R. A. (2001). Distributing working memory resources during problem solving. *J. Experi. Psychol.* 27, 836–848. doi: 10.1037/0278-7393.27.3.836

Cecutti, L., Chemero, A., and Lee, S. W. S. (2021). Technology may change cognition without necessarily harming it. *Nat. Hum. Behav.* 5, 973–975. doi: 10.1038/s41562-021-01162-0

Chien, S.-Y., Lewis, M., Sycara, K., Liu, J.-S., and Kumru, A. (2016). "Relation between trust attidudes toward automation, Hofstede's cultural dimensions, and Big Five personality traits," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. doi: 10.1177/1541931213601192

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., and Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Comput. Human Behav.* 127:107018. doi: 10.1016/j.chb.2021.107018

Craik, F. I. M. (2002). Levels of processing: Past, present...and future? *Memory* 10, 305–318. doi: 10.1080/09658210244000135

Craik, F. I. M., and Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *J. Verbal Learning Verbal Behav.* 11, 671–648. doi: 10.1016/S0022-5371(72)80001-X

Dawson, P. (2020). "Cognitive offloading and assessment," In: *Re-Imagining University Assessment in a Digital World*, eds M. Bearman, P. Dawson, R. Ajjawi, J. Tai, and D. Boud (Cham: Springer). doi: 10.1007/978-3-030-41956-1_4

Dunn, T. L., Gaspar, C., McLean, D., Koehler, D. J., and Risko, E. F. (2021). Distributed metacognition: Increased bias and deficits in metacognitive sensitivity when retrieving information form the internet. *Technol. Mind Behav.* 2:39. doi: 10.1037/tmb0000039

Eskritt, M., and Ma, S. (2014). Intentional forgetting: Note-taking as naturalistic example. *Mem. Cognit.* 42, 237–246. doi: 10.3758/s13421-013-0362-1

Fenech, E. P., Drews, F. A., and Bakdash, J. Z. (2010). "The effects of acoustic turn by turn navigation on wayfinding," In: *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*. doi: 10.1177/154193121005402305

Finley, J. R., Naaz, F., and Goh, F. W. (2018). *Memory and Technology: How We Use Information in the Brain and in the World*. Cham: Springer Nature Switzerland. doi: 10.1007/978-3-319-99169-6

Fisher, M., Goddu, M. K., and Keil, F. C. (2015). Searching for explanations: How the internet inflates estimates of internal knowledge. *J. Experi. Psychol.* 144, 674–678. doi: 10.1037/xge0000070

Fisher, M., Smiley, A. H., and Grillo, T. L. H. (2021). Information without knowledge: the effects of internet search on learning. *Memory*. 20, 375–387. doi: 10.1080/09658211.2021.1882501

Gilbert, S. J. (2015a). Strategic offloading of delayed intentions into the external environment. *Q. J. Exp. Psychol.* 68, 971–992. doi: 10.1080/17470218.2014.972963

Gilbert, S. J. (2015b). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Conscious. Cogn.* 33, 245–260. doi: 10.1016/j.concog.2015.01.006

Glikson, E., and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Acad. Manage. Ann.* 14:57. doi: 10.5465/annals.2018.0057

Gray, W. D., Sims, C. R., and Fu, W.-T. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychol. Rev.* 113, 461–482. doi: 10.1037/0033-295X.113.3.461

Grinschgl, S., Meyerhoff, H. S., and Papenmeier, F. (2020). Interface and interaction design: How mobile touch devices foster cognitive offloading. *Comput. Human Behav.* 108:106317. doi: 10.1016/j.chb.2020.106317

Grinschgl, S., Meyerhoff, H. S., Schwan, S., and Papenmeier, F. (2021a). From metacognitive beliefs to strategy selection: Does fake performance feedback influence cognitive offloading? *Psychol. Res.* 85, 2654–2666. doi: 10.1007/s00426-020-01435-9

Grinschgl, S., Papenmeier, F., and Meyerhoff, H. S. (2021b). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Q. J. Experi. Psychol.* 74, 1477–1496. doi: 10.1177/17470218211008060

Grinschgl, S., Papenmeier, F., and Meyerhoff, H. S. (2022). Mutual interplay between cognitive offloading and secondary task performance.

Hills, T., and Hertwig, R. (2011). Why aren't we smarter already: Evolutionary trade-offs and cognitive enhancements. *Curr. Dir. Psychol. Sci.* 20, 373–377. doi: 10.1177/0963721411418300

Hoff, K. A., and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–437. doi: 10.1177/0018720814547570

Hu, X., Luo, L., and Fleming, S. M. (2019). A role for metamemory in cognitive offloading. *Cognition* 193:104012. doi: 10.1016/j.cognition.2019.104012

idc.com (2022). *Smartphone Shipments Declined in the Fourth Quarter But 2021 Was Still a Growth Year With 5.7% Increase in Shipments According to IDC*. Available online at: https://www.idc.com/getdoc.jsp?containerId=prUS48830822

Jaušovec, N., and Pahor, A. (2017). *Increasing Intelligence*. London: Elsevier Academic Press.

Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Hum. Factors.* 47, 915–926. doi: 10.1177/00187208211013988

KMU Forschung Austria (2020). *E-Commerce-Studie Österreich 2020*. https://www.kmuforschung.ac.at/e-commerce-studie-2020/

Kraus, J., Scholz, D., and Baumann, M. (2021). What's driving me? Exploration and validation of a hierarchical personality model for trust in automated driving. *Hum. Fact.* 63, 1076–1104. doi: 10.1177/0018720822653

Li, W., Yao, N., Shi, Y., Nie, W., Zhang, Y., Li, X., et al. (2020). Personality openness predicts driver trust in automated driving. *Automotive Innovation* 3, 3–13. doi: 10.1007/s42154-019-00086-w

Liao, M. S., Sandberg, A., and Roache, R. (2012). Human engineering and climate change. *Ethics Policy Environ.* 15, 206–221. doi: 10.1080/21550085.2012.685574

Matthews, G., Hancock, P. A., Lin, J., Panganiban, A. R., Reinerman-Jones, L. E., Szalma, J. L., et al. (2021). Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Pers. Individ. Dif.* 169:e109969. doi: 10.1016/j.paid.2020.109969

Merrit, S. M., and Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Hum. Fact.* 50, 194–210. doi: 10.1518/001872008X288574

Meyerhoff, H. S., Grinschgl, S., Papenmeier, F., and Gilbert, S. J. (2021). Individual differences in cognitive offloading: a comparison of intention offloading, pattern copy, and short-term memory capacity. *Cognit. Res.* 6:34. doi: 10.1186/s41235-021-00298-x

Moritz, J., Meyerhoff, H. S., and Schwan, S. (2020). Control over spatial representation format enhances information extraction but prevents long-term learning. *J. Educ. Psychol.* 112, 148–165. doi: 10.1037/edu0000364

Neubauer, A. C. (2021). The future of intelligence research in the coming age of artificial intelligence–with special consideration of the philosophical movements of trans- and posthumanism. *Intelligence* 87:101563. doi: 10.1016/j.intell.2021.101563

O'Hara, K. P., and Payne, S. J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cogn. Psychol.* 35, 34–70. doi: 10.1006/cogp.1997.0676

Risko, E. F., and Gilbert, S. J. (2016). Cognitive offloading. *Trends Cogn. Sci.* 20, 676–688. doi: 10.1016/j.tics.2016.07.002

Runge, Y., Frings, C., and Tempel, T. (2019). Saving-enhanced performance: saving items after study boosts performance in subsequent cognitively demanding tasks. *Memory* 27, 1462–1467. doi: 10.1080/09658211.2019.1654520

Schäfer, K. E., Chen, J. Y. C., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Hum. Fact.* 58, 377–400. doi: 10.1177/0018720816634228

Schönpflug, W. (1986). The trade-off between internal and external information storage. *J. Mem. Lang.* 25, 657–675. doi: 10.1016/0749-596X(86)90042-2

Sharan, N. N., and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon* 6:e04572. doi: 10.1016/j.heliyon.2020.e04572

Siau, K., and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technol. J.* 31, 47–53.

Sorgner, S. L. (2020). *On Transhumanism*. Penn State Univeristy Press. doi: 10.5040/9781350090507.ch-003

Sparrow, B., Liu, J., and Wegner, D. (2011). Google effects on memory: Cognitive consequences of having information at out fingertips. *Science* 33, 776–778. doi: 10.1126/science.1207745

Spitzer, M. (2012). *Digitale Demenz: Wie wir uns und unsere Kinder um den Verstand bringen*. Droemer.

Storm, B. C., and Stone, S. M. (2015). Saving-enhanced memory: The benefits of saving on the learning and remembering of new information. *Psychol. Sci.* 26, 182–188. doi: 10.1177/0956797614559285

Van Nimwegen, C., and Van Oostendorp, H. (2009). The questionable impact of an assisting interface on performance in transfer situations. *Int. J. Ind. Ergon.* 39, 501–508. doi: 10.1016/j.ergon.2008.10.008

Waldron, S. M., Patrick, J., and Duggan, G. B. (2011). The influence of goal-state access cost on planning during problem solving. *Q. J. Experi. Psychol.* 64, 485–503. doi: 10.1080/17470218.2010.507276

Zhang, Z., Tao, D., Qu, Y., Zhang, X., Zeng, J., Zhu, H., et al. (2020). Automated vehicle acceptance in China: Social influence and initial trust are key determinants. *Transport. Res. Part C: Emerg. Technol.* 112, 220–233. doi: 10.1016/j.trc.2020.01.027

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 19 July 2022
doi: 10.3389/frai.2022.910630

On Distributed Cognition While Designing an AI System for Adapted Learning

Magne V. Aarset[1,2]* *and Leiv Kåre Johannessen*[2]

[1] Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology - NTNU, Ålesund, Norway, [2] TERP Research Department - TRD, TERP AS, Haugesund, Norway

When analyzing learning, focus has traditionally been on the teacher, but has in the recent decades slightly moved toward the learner. This is also reflected when supporting systems, both computer-based and more practical equipment, has been introduced. Seeing learning as an integration of both an internal psychological process of acquisition and elaboration, and an external interaction process between the learner and the rest of the learning environment though, we see the necessity of expanding the vision and taking on a more holistic view to include the whole learning environment. Specially, when introducing an AI (artificial intelligence) system for adapting the learning process to an individual learner through machine learning, this AI system should take into account both the learner and the other agents and artifacts being part of this extended learning system. This paper outlines some lessons learned in a process of developing an electronic textbook adapting to a single learner through machine learning, to the process of extracting input from and providing feedback both to the learner, the teacher, the learning institution, and the learning resources provider based on a XAI (explainable artificial intelligence) system while also taking into account characteristics with respect to the learner's peers.

Keywords: distributed cognition and learning, distributed situational awareness, adaptive learning, artificial intelligence, stochastic processes

OPEN ACCESS

Edited by:
Paul Seitlinger,
University of Vienna, Austria

Reviewed by:
Cali Michael Fidopiastis,
Design Interactive, United States
Birgitta Dresp-Langley,
Centre National de la Recherche Scientifique (CNRS), France

***Correspondence:**
Magne V. Aarset
magne.aarset@ntnu.no

Specialty section:
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

Received: *01 April 2022*
Accepted: *24 June 2022*
Published: *19 July 2022*

Citation:
Aarset MV and Johannessen LK (2022) On Distributed Cognition While Designing an AI System for Adapted Learning. Front. Artif. Intell. 5:910630. doi: 10.3389/frai.2022.910630

THE LEARNING SYSTEM

The Learning Environment

Ever since ancient times until today's society with all kinds of information readily available in an always present computer, (most) humans have understood that learning is vital. Socrates is supposed to have stated that:

- *The **wise** man learns from everything and everyone.*
- *The **ordinary** man learns from his experience.*
- *The **fool** knows everything better.*

Besides the obvious advice to keep our eyes and ears open, this quote also puts the learner into an environment that consists of more than the individual learner herself, and even more than a dual learner—teacher relationship. Still, the learning process has been seen as a tug of war between the learner and a teacher, where the responsibility for the learning outcome has moved from the learner to the teacher, and in recent decades slightly back to the learner. In education, the focus is now on *learning* as opposed to *teaching*, based upon the understanding of learning as a more active process for the learner than the more passive attitude that may be associated with teaching.

Learning needs to be seen as a process being executed in a much richer environment consisting of several agents and several cognitive artifacts (i.e., external representations of "knowledge of the world" as books, checklists, decision support systems, and language) (Norman, 1991).

Even though focus for ages has been a key word while studying processes involving human activities, science exploration today more and more sees the necessity of expanding the vision and take on a more holistic view as for example reflected in the works of Salomon (1997) and Hutchins (2001). Theoretical models have moved from describing the feeling, thinking, and acting of a single human being to incorporate as many as possible of the agents and artifacts that combined constitute a system working to accomplish a goal.

The unit of analysis here is the functional system consisting of a collection of agents and artifacts and their relations to each other. Aarset and Glomseth (2019) describe this as *integrated operations*, while Hollnagel and Woods (2005) introduce the term *joint cognitive systems*, where cognitive processes will occur and be distributed. Beside the learner and the teacher, both the learning institution, the learning resources provider (e.g., authors and/or publishers of textbooks), the peers for example in a class, and some more external agents like employer, colleagues, family, friends, and sometimes even a community who have invested in "the smartest kid in the village" will all be part of this environment.

What's typical with such integrated operations are that the different participants may have different background, both regarding knowledge of what's supposed to happen and experience from similar operations, different individual goals, and finally, sometimes surprisingly different understanding of what's really going on. Such differences in goals, attention, perception, and roles to play are of course just as it should be, but insufficient understanding of what's really really going on, i.e., acquired and maintained *situational awareness* (Salmon et al., 2009), may cause actions performed with the best intentions that have adverse effect.

Therefore, it's necessary to have a system perspective, and incorporate the social interactions between the agents, the interaction between the agents and the artifacts, and the means of organizing this into a productive unit. We need to identify the components within the system and explain the mechanisms that coordinate this group of collaborators.

Objectives

The objective of the cognitive learning system we are analyzing here is for one single learner to learn. That is, we don't see this as a group who is supposed to learn to collaborate while executing some process later, as for example a crew operating an airplane or a ship. For the learning process of one such single learner the overall objective may be threefold. That is to increase the learner's

- competence,
- confidence,
- learning ability.

Increasing the learner's competence may according to the revised Bloom's taxonomy for knowledge-based learning (Bloom, 1956 and Anderson and Krathwohl, 2001) be to enable the learner to both

- *remember*; find or remember information,
- *understand*; understanding and making sense of information,
- *apply*; use information in a new, but similar, situation,
- *analyze*; take information apart and explore relationships,
- *evaluate*; critically examining information and make judgments,
- *create*; use information to create something new.

Furthermore, it is a goal in itself to get a learner to be confident enough to apply what has been learned. It's of little use to present something worth knowing to learners if they don't feel confident enough to act according to it.

Finally, in a world that is constantly changing, we might say that there is no single subject or set of subjects that will serve a learner for the foreseeable future. The most important skill to acquire may be learning how to learn. Therefore, it is beneficial to improve the learner's situational awareness with respect to her own learning ability by for example giving feedback on how the learner is utilizing the learning resources, compared to her peers.

During learning typical goals of typical agents in a learning system may for example be like illustrated in **Figure 1**.

In addition to the different individual goals each agent should operate within some constraints, which typically may be the time and resources available, and the well-being of the agents in the learning environment.

Process Flow

To illustrate the system design and the overall flow during an integrated operation, a practical and convenient technique is to use SADT diagrams (Marca and McGowan, 1988). SADT sheets are a combination of activity boxes and arrows indicating the order in which the activities are to be carried out. An ICOM system (Input, Control, Output, Mechanism) is distinguishing between

- *Input* or input data from the left of the activity box, which is something that should be changed by or starting the activity.
- *Output*, which is the result of the activity.
- *Control*, which decides when and how the activity it to be performed (typically within some constraints).
- *Mechanism*, which is identifying the agents and the artifacts that performs the activity.

In the simplified example in **Figure 2**, four phases have been identified. First, a "pre-learning phase" (*preparation*) that may influence the learners' "starting competence" and/or motivation. The *participating* phase is where the learner may be in a learning environment with (several) other agents, or quite alone with some learning resources. Hopefully, the learner will use some time for reflection in the *pondering* phase, before using her new knowledge in a practical situation.

To illustrate what's ideally going on in the *participating phase* we may lean on the four stages identified by Kolb (1984) for achieving effective learning. According to Kolb a learner should progress through a cycle of the four stages:

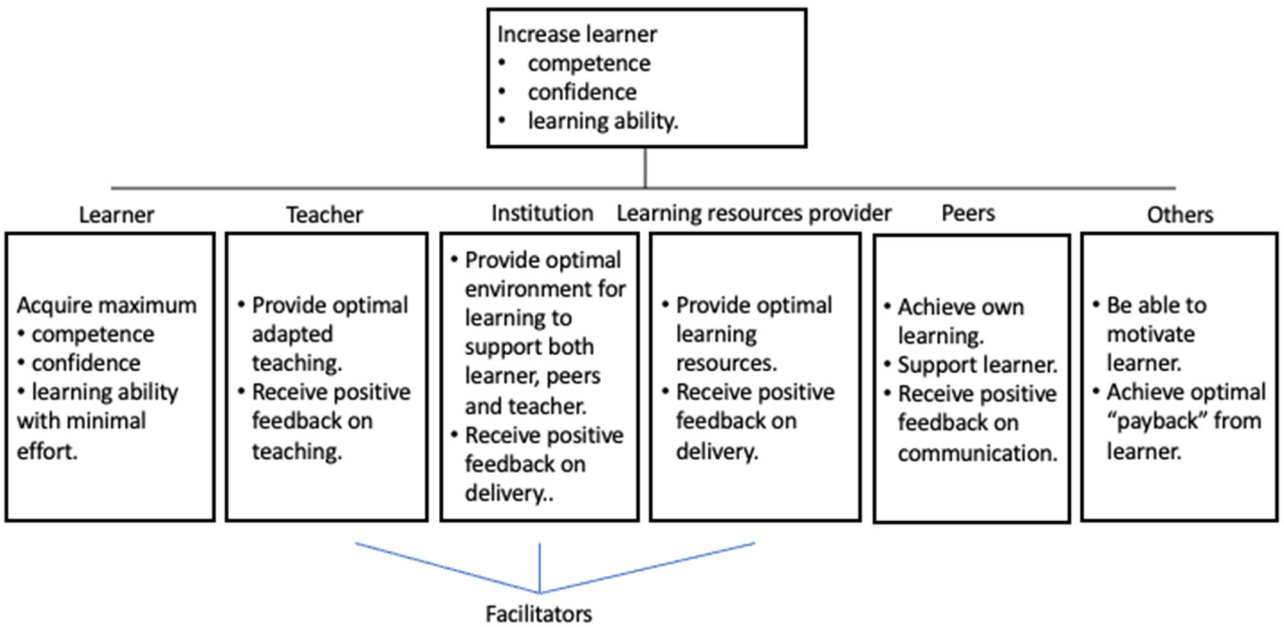

FIGURE 1 | Objective hierarchy during learning.

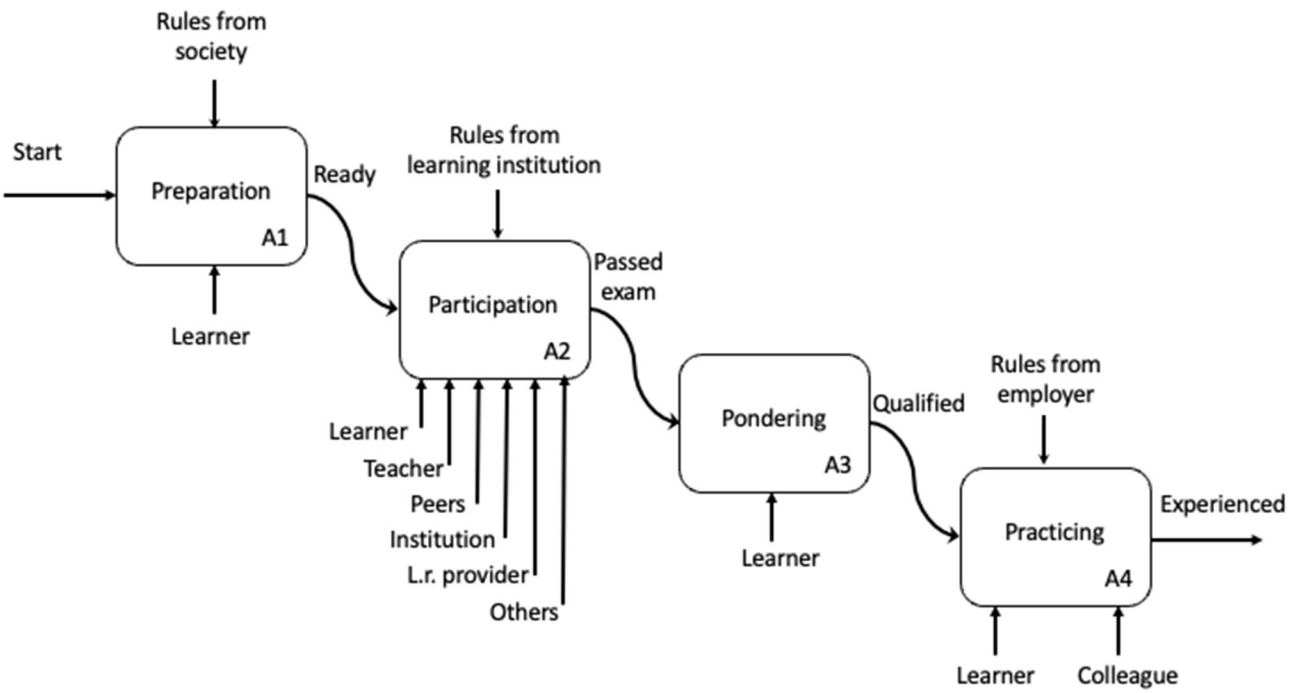

FIGURE 2 | Illustration of the different phases a learner goes through.

- having a *concrete experience*,
- having an *observation* of and *reflection* on that experience,
- forming *abstract concepts* (analysis) and *generalizations* (conclusions),
- testing *hypotheses* in future situations, resulting in new *experiences*.

Interactions

None of the above illustrations or models are well suited to illustrate the communication between the different agents and/or the artifacts, though. These relations may be illustrated by introducing so-called *agent-based flow charts* (ABFC) (Aarset, 2014; Aarset and Glomseth, 2019).

As the name indicates, the emphasis here is on visualizing the connection between the different agents, between the different agents and the artifacts, and how they relate to each other. Admittedly, when using such agent-based flow charts it is often more difficult to see how an activity is performed from the beginning to the end, but it is significantly easier to see what information each agent needs to be able to perform her activities (functions), and what information and what result each agent should pass on. It is also easier to see and understand when and to which other agent(s) this result should be passed on to.

The agents and artifacts that make up the system are identified from the mechanism inputs in the SADT sheets. There should be constructed one agent-based flow chart for each agent (and sometimes also for some artifacts). Each such agent-based flow chart is constructed by listing all the activities that shall be performed by this agent in a box placed in the middle of the chart, e.g., standard operating procedures (SOP). Then identify for each of these activities separately whether the agent who is going to execute these activities needs input (information or commands) from another agent/artifact. These inputs are illustrated by drawing arrows from smaller boxes from each of the relevant other agents/artifacts to the left of the main box. For each such "reporting agent," identify which input data she will transfer to the agent in focus and which activity to be executed by this "reporting agent" this "reporting" is related to.

Finally, boxes are created on the right side of the main box for those of the other agents (or artifacts) that are to receive something from the agent in focus. A schematic illustration of a part of one agent-based flow chart, where the learner is the focus agent, is shown in **Figure 3**.

As each input to each focus agent per definition also is an output from another agent (or artifact), these flowcharts are particularly useful when checking that all agents are aware of their responsibility of what and to whom they are supposed to report. Observe also that this is an internal analysis. No external input from outside the learning environment, or output to this external environment, are considered.

Distributed Situational Awareness

Still another way of understanding learning in a rich learning environment is to follow the logic of Salmon et al. (2009) with respect to distributed situational awareness. They view distributed situational awareness as "the system's collective knowledge regarding a situation that comprises each element's compatible awareness of that situation." Their model (**Figure 4**) uses schema theory and Neisser's perceptual cycle model (Neisser, 1976) with respect to each agent and treats distributed situational awareness as "a systemic property that emerges from the interaction (referred to as situational assessment transactions) between system elements (human and non-human)" (Salmon et al., 2009).

When performing an integrated operation as learning in a rich learning environment, Salmon et al. (2009) will classify the activities to be carried out by the involved agents as either *teamwork* or *taskwork*. Teamwork is activities where the behavior of the actors is affecting each other, or they coordinate their behavior in relation to each other. Taskwork means activities where the actors are performing individual activities separately and (in part) independently of input from the other actors to reach the system's partial or overall objective.

We see that the models (illustrations) in **Figures 1–3** above give information which directly may be included in this model. **Figure 1** gives input to the *System goals* in the *System factors*, and to *Goals and roles* in *Individual factors*. **Figures 2**, **3** give input to *System design* and *Procedures* (also in the *System factors*), while Bloom's taxonomy provides input to both *Task factors*, *Team factors*, *Individual factors* and *System factors*.

LEARNING

Modeling Learning

In line with Illeris (2009) learning may be understood as *a process that leads to a permanent capacity to change which is not solely due to biological maturation or aging*, and that the learner during learning constructs mental structures (*schemes*) processed within the memory function (see e.g., Piaget, 1973; Neisser, 1976; Vygotsky, 1978). This process implies both the integration of an *external interaction process* between the learner and the other agents and artifacts in the learning environment, and an *internal psychological process* of acquisition and elaboration.

Furthermore, this *internal psychological process* is a process of integrated interplay between a content dimension (*competence*) which concerns both what is to be learned and the learner's abilities (understanding, knowledge, skills, etc.), and an incentive dimension (*commitment*) which provides and directs the mental energy that is necessary for learning to take place (motivation, emotion, volition, etc.) (**Figure 5**).

We prefer the headings *competence* and *commitment* instead of Illeris' terms *content* and *incentive* partly to be in line with the terms from Situational Leadership Theory (SLT) (Thompson and Aarset, 2012), which will be utilized as a basis for the feedback approach.

To transform this model into a mathematical/statistical model of the learning process making it possible to characterize, evaluate, and adapt to an individual learner autonomously, all the above (or similar) suggested models and techniques illustrated in **Figures 1–4** are necessary steps. Such a mathematical/statistical model may form the basis for utilizing technology to improve the learning process by giving feedback. It is convenient to illustrate such a mathematical/statistical model by conceptual diagrams.

Conceptual diagrams illustrate a set of relationships between variables (Hayes, 2018). An antecedent variable X may in addition to a direct effect on a consequent variable Y also cause variation in one (or more) *mediator variable(s)* M1, which, in turn, also causes variation in Y (see **Figure 6**). Here, a typical example of a mediator variable is motivation. The available learning resources are for example directly influencing the learning outcome. Still, they may also influence the learner's motivation and are therefore in addition influencing the learning outcome indirectly.

Furthermore, the association between two variables X and Y is said to be *moderated* when the effect of an antecedent variable X on a consequent variable Y depends on a third variable (or set of variables) M2. Here, a typical example of a moderator

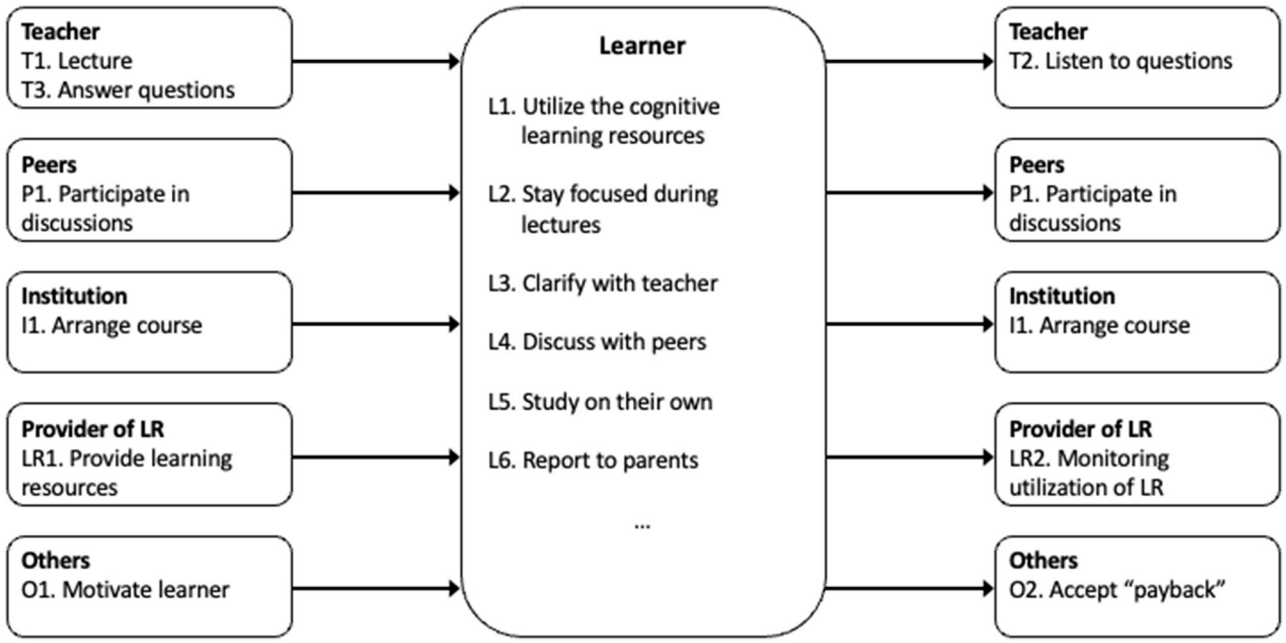

FIGURE 3 | A simplified agent-based flowchart focusing on the learner.

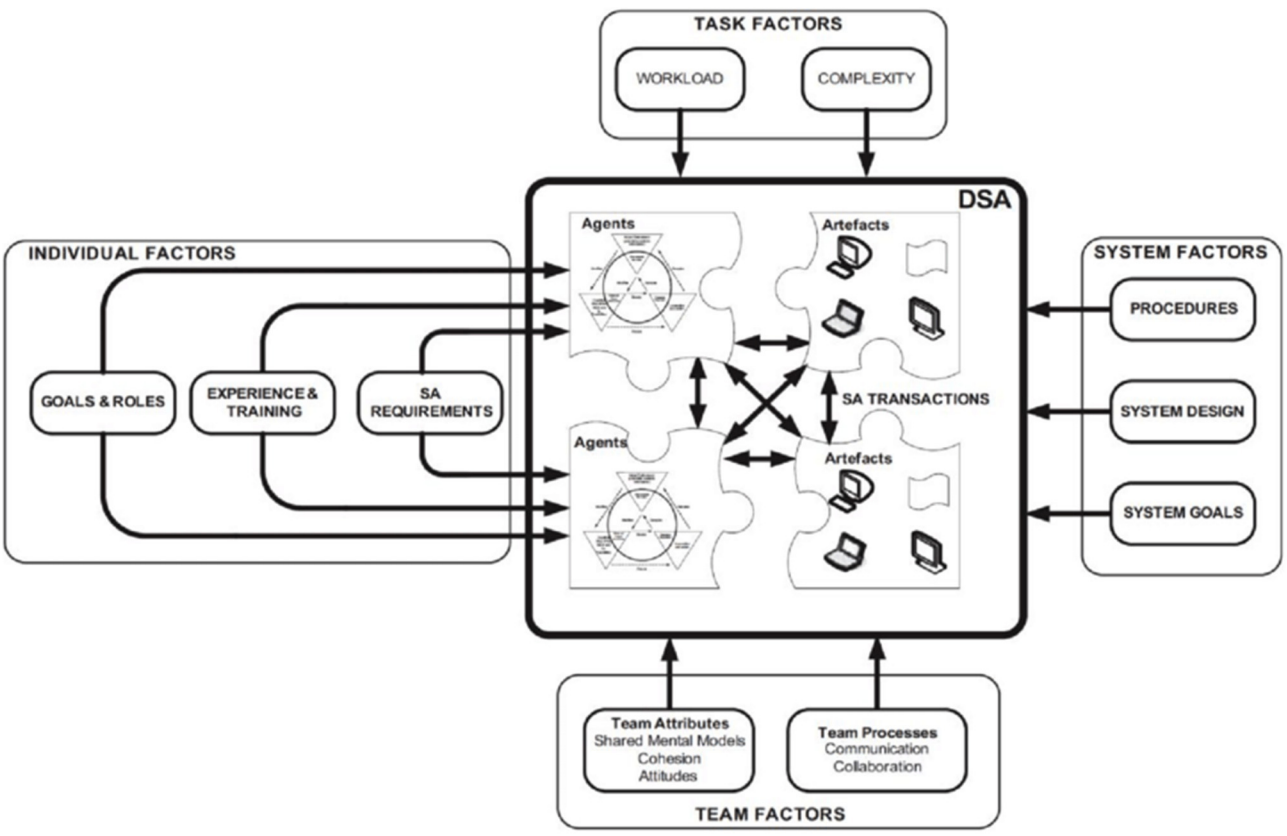

FIGURE 4 | Illustration of distributed situational awareness (Salmon et al., 2009).

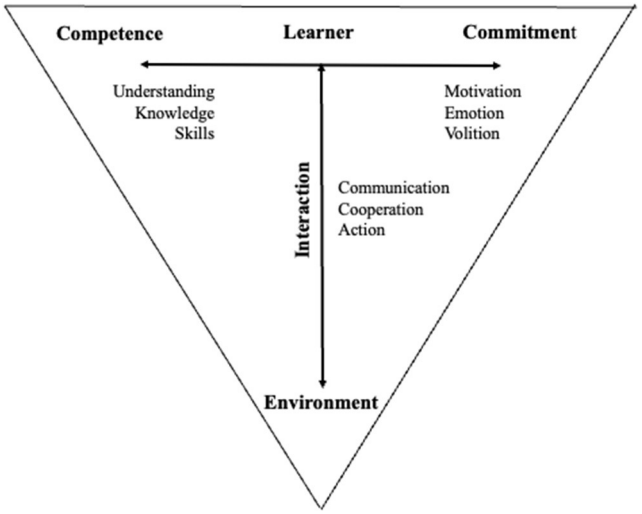

FIGURE 5 | An illustration of the three dimensions of learning [inspired by Illeris (2009)].

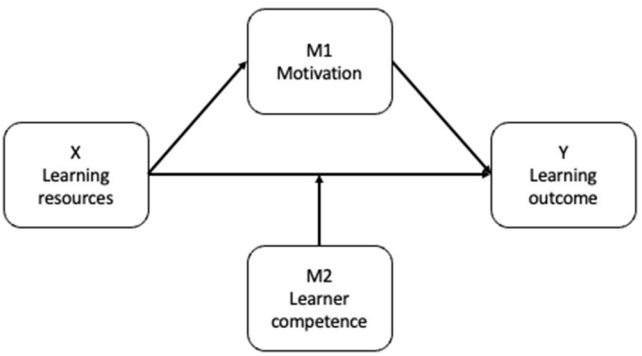

FIGURE 6 | Conceptual diagram with mediator (M1) and moderator (M2) variables.

variable is competence. It may for example be assumed that how the available learning resources influence the learning outcome depends on the learner's initial level of competence and ability to acquire knowledge.

The conceptual diagram in **Figure 7** below illustrates the activities in a limited time frame of a learning process, let's call it *a learning session*, where we assume that only the characteristics competence, confidence, and learning ability change during this time interval. Feedback to the different agents, which may lead to a change of state or activity, will only be presented at the end of such learning sessions.

The Learner

As stated in the objective hierarchy in **Figure 1**, the development of the learner's level of competence, confidence, and learning ability are the key factors the learning process is intended to improve. Factors included in the model in **Figure 7** with respect to the learner contains in addition both "*telemetry*" (TM) such as sociocultural, sociodemographic, and socioeconomic factors, personality (e.g., according to McCrae, 2018), as well as the objective of the learner. It is assumed that these additional factors don't change during a learning session, and that there is a direct effect of all these factors on the learning outcome.

All these factors are also assumed to provide an indirect effect on the new competence, new confidence, and new learning ability through the influence on the commitment, the chosen learning approach, and the learner's utilization of the learning environment. Furthermore, it is assumed that these factors will moderate the effect of the learning resources on the learning pathway itself.

Influence From the Other Agents

Influence from the other agents in the learning system will take different forms. Both the teacher and the peers are assumed to influence the learner's commitment, chosen learning approach, and utilization of the learning environment. They are also both expected to moderate the learning pathway.

The same is expected to hold for the learning resources, while the learning institution is expected to influence the learner for example through their organization of a study program at a university, or an internal course in a company. The other agents are assumed only to influence the learner's commitment.

Commitment

The commitment of the learner is assumed to be an important factor with respect to the learning outcome. Herzberg (1982) suggests that motivational factors may be split in two groups.

- *Motivators* that give positive satisfaction, arising from intrinsic conditions of the learning process itself (e.g., personal growth, opportunity to do something meaningful, sense of importance).
- *Hygiene factors* that do not give positive satisfaction or lead to higher motivation, just dissatisfaction in case of their absence (e.g., status, work conditions, vacations).

Both emotion (Um et al., 2012) and volition (Garcia et al., 1998) are known to have a direct effect on learning and will thereby also affect the learning outcome. Um et al. (2012) conclude that induced positive emotions in learners both will enhance comprehension of content and facilitate the construction of mental models required for utilization of information in a new, but similar, situation.

Volitional processes are defined as those thoughts and behaviors that are directed toward maintaining one's intention to attain a specific goal in the face of both internal and external distractions (Corno and Kanfer, 1993). Beside encoding information into the long term memory store the instrumental strategies involved during learning also include volitional strategies to maintain the intention and the attempts to learn. According to Corno (1993), volition plays a mediating role between the intention to learn and the use of learning strategies.

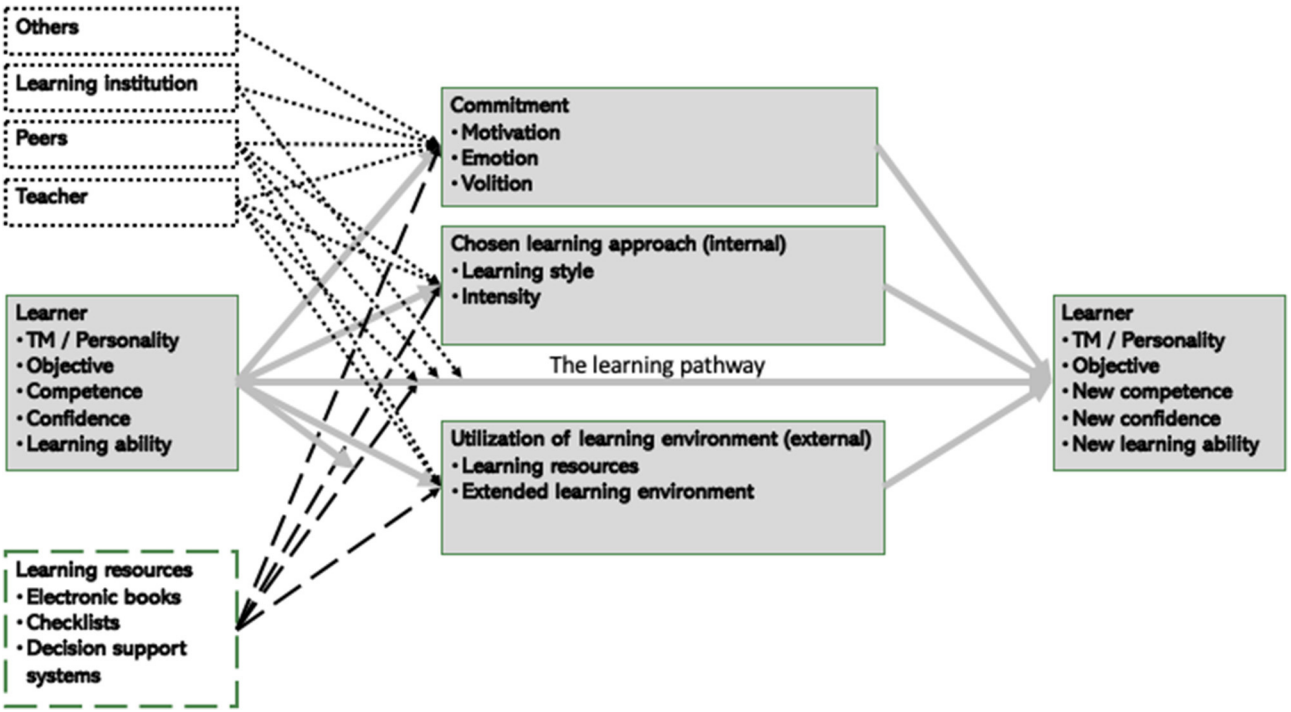

FIGURE 7 | Learning from the perspective of the learner.

Chosen Learning Approach

The chosen learning approach taken by the learner is also assumed to influence the learning outcome. Sternberg (1994) suggests that learning styles can be understood in terms of functions, forms, levels, scope, and leanings of government.
Functions:

- Legislative; Define objective and plan strategy.
- Executive; Execute predefined strategies.
- Judicial; Evaluate/criticize objectives and/or strategies.

Forms:

- Monarchic; Direct focus on one goal at a time.
- Hierarchic; Sees whole picture and prioritize.
- Oligarchic; Sees whole picture, but doesn't prioritize.
- Anarchic; Sees whole picture, but selects a random approach.
- Democratic; Sees whole picture, and pleases everyone.

Levels:

- Local; Bottom-up.
- Global; Top-Down.

Scope:

- Introvert; During execution.
- Extravert; During execution.

Leanings:

- Liberal; Openminded, Prefer changes.
- Conservative; Sticking to established rules.

Sternberg's classification is debated in the literature, though, but is still a useful starting point when searching for proxy variables to be included in the mathematical/statistical AI model.

An alternative way of studying chosen learning approach is to distinguish between the strategies rehearsal, elaboration, and organization (Garcia et al., 1998). *Rehearsal* strategies are used to select and encode information in a verbatim manner (e.g., repetition of information). *Elaboration* strategies are used to make information meaningful and to build connections between information given in the learning assignment and a learner's prior knowledge (e.g., mental imagery, use of mnemonics, creating analogies, and trying to teach the information to another person). *Organizational* strategies are used to construct internal connections among the pieces of information to be learned (e.g., clustering related information based on common characteristics).

Furthermore, the intensity of how the learner is acting is also assumed to affect the learning outcome.

Utilization of the Learning Environment

As long as one human mentor to each learner at all times are probably neither possible nor desirable, the introduction of an AI system may be a suitable alternative option. Beside this, (electronic) learning resources hold several opportunities to enhance both motivation and learning. This may be through an adaptive electronic textbook to the individual learner, which both may give the opportunity for communication within the learning environment, and also to include *immersive environments* to facilitate better, deeper learning.

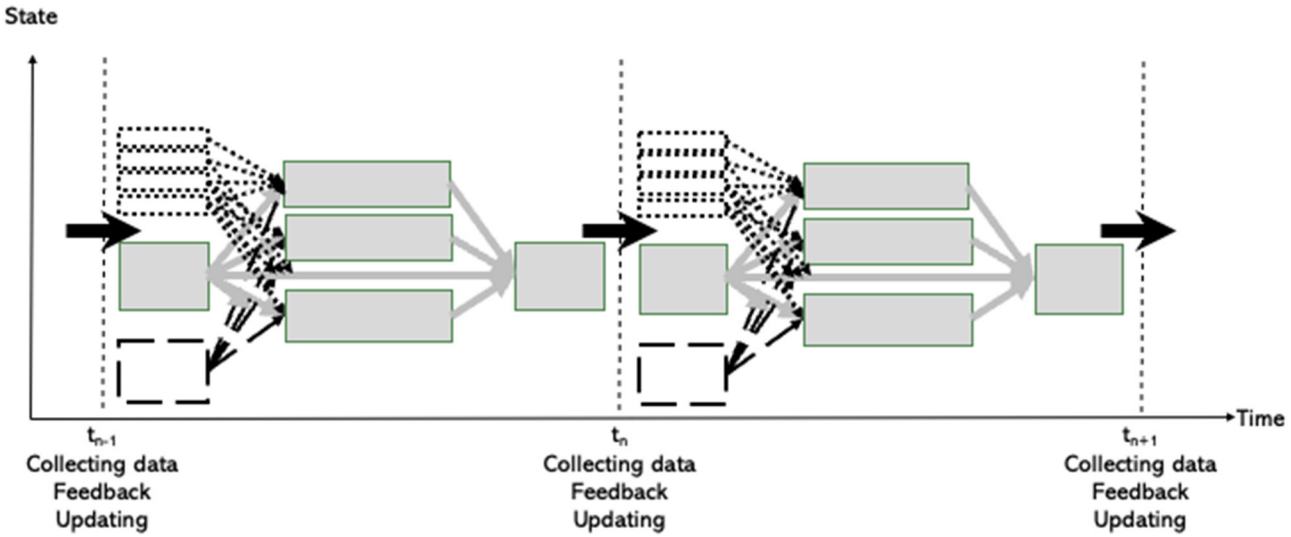

FIGURE 8 | The learning process evolving in time.

Giving a student the opportunity both to see a newly presented detailed explanation of some concept into a larger whole, and maybe even to observe consequences after experimenting with this larger understanding, are desirable. Augmented reality can do this through enhanced natural environments or situations that offer perceptually enriched experiences.

It's common to distinguish between three types of immersive interfaces (Dede et al., 2017).

- *Virtual Reality* (VR) interfaces provide exclusive input to our senses as response to our actions to simulate a real world setting.
- *Multiuser Virtual Environments* (MUVE) interfaces provide input from a virtual environment to digital avatars.
- *Mixed Reality* (MR) combine real and virtual settings, for example by superimposing information (*Augmented Reality*, AR) onto the view of a real world setting.

All three capabilities may improve learning by simulating that learning takes place in a similar context to that in which it is later supposed to be applied (*situated learning*). How the learner is utilizing both such opportunities alone and in collaboration with a teacher and/or peers, may be important with respect to the learning outcome.

Updated State of the Learner

At the end of such a learning session as described here, it is assumed that the "telemetry" (TM), the personality of the learner and the learner's objective, are unchanged, but that the learner has reached a new level of competence, confidence, and learning ability.

THE LEARNING PROCESS

The Model

The learning process may be understood as a discrete time stochastic process (hopefully with positive drift). That is, a family $\{\mathbf{X}_t : t \in T\}$ of random vectors $\mathbf{X}_t$, indexed by some set T, where each random vector will take on values from the same state space characterizing the states of each of the agents and artifacts involved in the learning process. The focus should primarily be on the learning pathway, that is, how the competence, confidence, and learning ability of the learner is developing, in conjunction with the state of the rich extended learning environment. The AI system providing feedback into the learning environment based on the system state is itself a part of this learning environment.

Let's first suggest a model for the learning process that might be valid for a shorter time period, what we earlier called *a learning session*, and describe the state space of this system (i.e., an identification of who and what is included in the model and a characterization of each of the agents and artifacts) at the beginning of time t_{n-1} and at the end of this time period at time t_n. This may be illustrated in the conceptual diagram in **Figure 7**.

The initial state of the system at time t_{n-1} will develop into a new state at time t_n. Then, at time t_n, an AI system will give feedback into the learning system. The state of the system will be revised simultaneously at this time t_n, and constitute the initial state used as input to the next time interval starting at time t_n. The model within each learning session will be the same, but at the end of each learning session the AI system will provide some feedback into the learning system and the values of the random vectors $\mathbf{X}_t$ are regularly being updated, as illustrated in **Figure 8**.

The state vector may be on the form as

$$\mathbf{X}_{t+1} = (\mathrm{TM}, \mathrm{P}, \mathrm{O}, \mathrm{Com}_t, \mathrm{Con}_t, \mathrm{LA}_t, \mathrm{Os}, \mathrm{LI}, \mathrm{Ps}, \mathrm{Te}, \mathrm{LR}, \mathrm{C}, \mathrm{CLA}, \mathrm{ULE}, \mathrm{Com}_{t+1}, \mathrm{Con}_{t+1}, \mathrm{LA}_{t+1}, \mathrm{FB}_{t+1})$$

where

- TM: Characteristics of the learner.
- P: Personality of the learner.
- O: The objective of the learner.
- Com_t: The competence of the learner at time t.

- Con_t: The confidence of the learner at time t.
- LA_t: The learning ability of the learner at time t.
- Os: Information of "the others."
- LI: Information from the learning institution.
- Ps: Information regarding the peers.
- Te: Information regarding the teacher.
- LR: Information regarding the available learning resources.
- C: The commitment of the learner.
- CLA: The chosen learning style of the learner.
- ULE: The learner's utilization of the learning environment.
- Com_{t+1}: The competence of the learner at time t + 1.
- Con_{t+1}: The confidence of the learner at time t + 1.
- LA_{t+1}: The learning ability of the learner at time t + 1.
- FB_{t+1}: The feedback from the AI system at time t + 1.

The state vector must for all practical purposes be modeled as a *Markov process* (Cox and Miller, 1987), but may include more historical observations than just from one earlier time period. Each of these elements of the state vector will be a vector itself. The identification of significant (and available and measurable) attributes, with respective metrics, will obviously be difficult, but such a model may be seen as a partly ideal theoretical description suitable as a starting point for collecting significant data.

The Observations

Generally, the goal of mathematical/statistical models are to facilitate

- *describing* what's going on,
- *understanding* the causes of what's going on,
- *predicting* what's going to happen,
- *influencing* through controlling the causes.

This requires valid, reliable, and significant measurements through either stated or reveled preference. Therefore, the measurements should ideally be

- *operational, valid, and reliable*; they should with a certain level of precision measure what they are supposed to measure,
- *complete*; they should cover most of the important aspects of the objective,
- *minimal*; the problem should be kept as simple as possible,
- *measurable*; it should be possible to assign both a probability of the different possible outcomes and a preference between these possibilities.

Wishing for a complete set of measurements without including the richer learning environment than the learner—teacher duo seems in vain. To have the opportunity to include all significant information during a learning process is on the other hand creating some undesirable secondary effects, especially with respect to personal security. The protection of personal data will introduce issues that must be handled satisfactory, both from a legal perspective and also reflecting what kind of information a learner may find it acceptable to share. Generally, this should be covered when acting according to the General Data Protection Regulation (GDPR) and the Data Protection Law Enforcement Directive.

Information to be included in the mathematical/statistical model with respect to the learner will vary over time and will for all practical purposes basically contain *proxy variables*, measurements just reflecting the real characteristics (Aarset, 2014). Internal attributes as assignment marks, quizzes, attendance, cumulative grade point average, etc. may easily be registered and utilized, while some external attributes as extra-curricular activities, social interaction network, personal interest, study habits, family support, etc., may be harder both to measure, to get access to, and to utilize within sound ethical constraints as stated in the general data protection regulations (GDPR).

Communication through the canals available in the digital learning system with a teacher may for example be expected to be more directly on the subject and for some learners relatively frequent. Direct communication with the other agents may be less frequent, but on the other hand maybe more continuously present in the mind of the learner. Available form of communication between the learner and the teacher, between the learner and the peers, and continuously updated reveled time and form of this kind of communication should be registered. The resources provided by the learning institution should also be registered and included in the mathematical model, as some characteristics of the other groups of agents.

A realization of such a (stochastic) learning process will provide data from each learning session, basically based on learner activity. A part of the data characterizing learner activity may for example be as illustrated in **Figure 9**. Here we see a learner who has started reading before watching a video, and then reading again before an idle period. After the break the learner is watching videos and an animation before taking an assessment.

Such activities occurring in the learning sessions will be repeated several times during the realization of the learning process as illustrated in **Figure 10**.

Adaptive learning is thus seen as a repeated process of collecting data from the learning system, utilizing these data for understanding the learner's progress, and then repeatedly providing feedback back into the learning system. Therefore, the data collected from the learner activity must be augmented with more data from the learning environment.

It is difficult to measure improvement in both competence, commitment, and learning ability. Let's for example assume that we despite this difficulty choose to measure the level of competence by the score on an assessment. Even though it may be realistic to model this as a stochastic variable, it is for example not at all clear which probability distribution we would prefer of the respective probability distributions illustrated in **Figure 11**.
Solid line probability density:

- Expected score = 60%.
- Probability of "high score (>75%)" ≈ 0.
- Probability of "low score (<50%)" ≈ 0.

Dashed line probability density:

- Expected score = 63%.
- Probability of "high score (>75%)" ≈ 0.1.
- Probability of "low score (<50%)" ≈ 0.1.

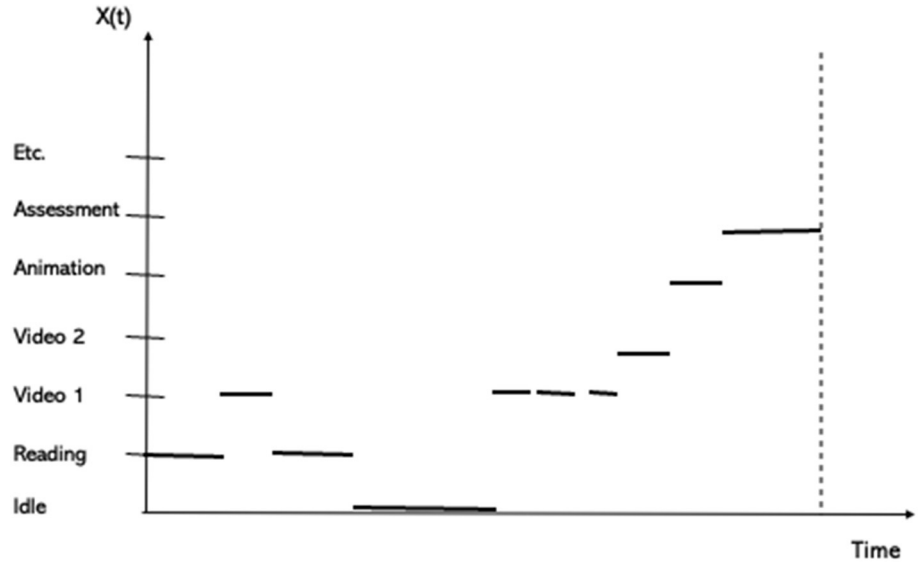

FIGURE 9 | An illustration of some of the activities of a learner during a learning session.

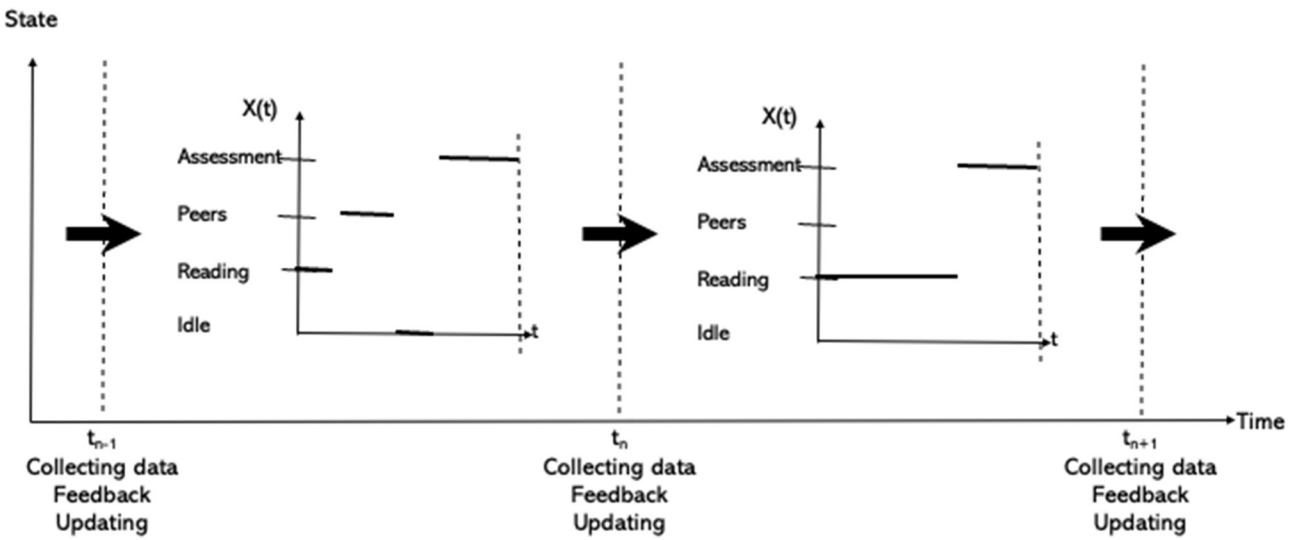

FIGURE 10 | An illustration of some of the activities of a learner during two learning sessions.

Dotted line probability density:

- Expected score = 64%.
- Probability of "high score (>75%)" ≈ 0.3.
- Probability of "low score (<50%)" ≈ 0.2.

There are similar issues with respect to other characteristics.

Censoring

Most learning processes will produce censored data. Learners who feel they don't have a satisfactory understanding of a subject may typically fail to register for a respective exam. A class at a university may for example have an improved average grading compared to last year's class, but with fewer students signed up for the exam. Not taking this censoring into account may reward an undesirable pedagogical approach (and the theoretical model would produce biased estimates).

The type of censoring most commonly seen when assessing knowledge is a sampling procedure where we only observe an assessment T_i if $T_i > C_i$ ($i = 1, \ldots, n$). Generally, $C_1, \ldots, C_n$ are assumed to be mutually independent stochastic variables independent of $T_1, \ldots, T_n$, indicating at which knowledge level the learners themselves feel they need to be at before registering for a test or exam. That is, each learner is evaluating herself before deciding to do an assessment or not. If they feel they don't have

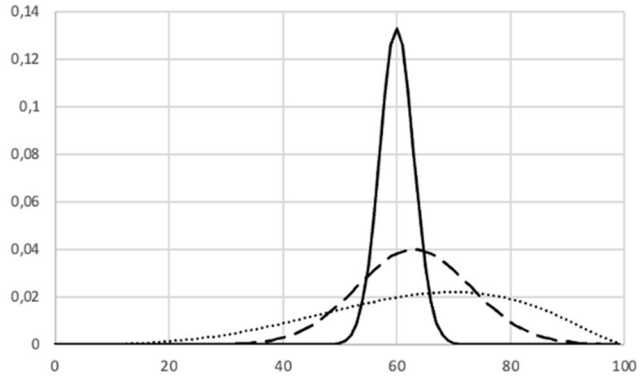

FIGURE 11 | Three probability densities representing a proxy variable for learning.

enough knowledge or understanding, some will abstain from taking the test.

ARTIFICIAL INTELLIGENCE

Machine Learning

To repeatedly and almost continuously produce adaptive feedback as decision support into a learning process may require more resources than most learners have available. With the scientific advancements of available big data and artificial intelligence, though, several decision-makers today are increasingly relying on machine learning to provide feedback as decision support. Therefore, mathematical/statistical models embedded in AI systems are introduced into the learning environment, where AI is defined as systems performing actions, physical or digital, based on structured or unstructured data, for the purpose of achieving a given goal. Now is the time also to introduce such systems to improve the learning process.

Utilizing AI also makes it possible to acquire information "hidden" in the realizations of these stochastic processes. For example, to group learners requiring similar adapted support into clusters. This information provides input to the autonomous decision support system which in turn provides feedback both to the learners, the teachers, the educational institutions, and the learning resource providers.

Cluster Analysis

Cluster analysis is the art of finding groups in data (Kaufman and Rousseeuw, 1990) and has become a popular technique within unsupervised learning as a part of machine learning (Murphy, 2012). Let $O = \{o_1, o_2, \ldots, o_m\}$ be a set of "objects" (here learners). A partition divides O into subsets (clusters) $O = \{O_1, O_2, \ldots, O_k\}$ that satisfy $O_i \cap O_j = \Phi$ $(\forall\ i \neq\ j)$ and $O_1 \cup O_2 \ldots \cup O_k = O$. The objective is to find groups in such a way that objects in the same group are similar, while objects in different groups are as dissimilar as possible (**Figure 12**). Here, the different learners should be grouped into clusters where all members of a cluster will benefit from the same didactic technique.

Before any meaningful computation can be performed as part of such unsupervised learning, though, human intervention is called for in the following four steps;

1. selecting the *attributes* to characterize system states (i.e., the agents, the agent's behavior, and the artifacts),
2. selecting suitable *metrics* to quantify the selected attributes,
3. defining so-called *dissimilarities* to measure the distance between objects, objects and clusters, and between clusters,
4. selecting an *algorithm* to create the clusters.

The actual choice made in each of these steps will influence the final classification and thereby the reliability and validity of any decision support system. In many applied analyses, however, surprisingly little attention has been put on steps 1–3.

The technique of K-medoids cluster analysis can identify clusters in the multidimensional space spanned by characteristics of the learner, observations of the learner's utilization of the learning environment, the learning environment itself, and, specially, utilization of the learning resources. The goal is to automatically detect patterns in data and using the uncovered patterns to predict future outcomes of interest.

Suppose there are m learners to be clustered by means of F characteristics as indicated in Chapter The Model above (an augmented vector characterizing a learner, the utilization of the learning resources, and the status of the rest of the learning environment). Then, the data will be on the form of attributes (an F-dimensional vector) for each object, so that the measurements can be arranged in an $(m \times F)$ matrix, where the rows correspond to the objects and the columns to the different variables.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1F} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mF} \end{bmatrix}$$

This clustering will give input to the forecasting of the learning process, which again will form the basis for feedback into the learning system.

FEEDBACK

Introduction

Optimizing the distributed cognition in the joint cognitive learning system requires feedback both to the learner and to the extended learning environment. In accordance with the conceptual model presented in **Figure 7** above this feedback should cover aspects with respect to both competence, confidence, learning ability, and motivation of the learner as well as a description of the state of the system itself. Presenting feedback that is effective and appropriate at the right time to the right agent is key for the success of an AI system and should be adapted to the respective receiver.

Introducing new technology such as an AI system into the learning environment may in addition to improve learning bring about behavioral changes. The different *positions* of the agents won't change, but the *roles*, i.e., what people in these positions do and how they do it, and the *role relationships*, i.e., with whom they interact or how they interact, may change (Barley, 2020).

An AI system for decision support may not just transform what it means to be a student, but also what it means to be a teacher. The cultural expectations about how, when, where, with what, and with whom the role should be played may change. The AI system should both attend and give feedback on the interaction order, i.e., how the situated, patterned, and recurrent ways of behaving and interacting that mark a particular context are developing.

XAI—eXplainable Artificial Intelligence

Experience with human behavior tells us that it is not at all clear that a learner (nor a teacher, a learning resource developer, a learning institution, etc.) necessarily will follow advice they don't understand. Therefore, to be successful, such a decision support system providing feedback to the learning system will need to be based on what has been named XAI (*eXplainable Artificial Intelligence*) (Arrieta et al., 2020).

Explainable Artificial Intelligence is artificial intelligence where the feedback from the autonomous system, and the reasoning behind this feedback, can be understood and meaningfully be evaluated by humans. This is in contrasts to the concept of the "*black box*" principle, where even the system designers not necessarily can explain why an AI algorithm arrives at a specific result. Therefore, it's both beneficial and necessary to present results in a "*white box*" setting for improving the distributed situational awareness.

Such XAI systems will usually produce a large amount of data. It's easy, though, even for an AI system, to become "overconfident" with an abundance of observations and almost "require," instead of suggesting, a change in behavior. It is important to remember that even when an apparently massive data set is available for analysis, the effective number of data points for several important cases of interest might be quite small. So, in what probably also are the words of Socrates: *Few things are common. Most things are quite rare.*

Feedback to Acquire and Maintain Situational Leadership

To be able to meaningfully evaluate the feedback from an autonomous AI system, and the reasoning behind this feedback, the agents need to acquire and maintain a satisfactory level of situational awareness. Their situational awareness will influence their attention and control how they act. Therefore, the feedback from the AI system must describe the state of the system to facilitate this acquisition and maintenance and be in accordance with the model described in **Figure 4**.

The Form of the Feedback to the Learner

Suggestions of the form of the feedback to the learner may be based on the situational approach of leadership developed by Hersey and Blanchard (1969). The premise of their theory is that different development level of a follower, here a learner, requires different kind of leadership, here feedback from the AI system. Leaning on this theory the feedback to the learner should either be *directive* or *supportive*, depending on the learner's *competence* and *commitment*, i.e., *development level*.

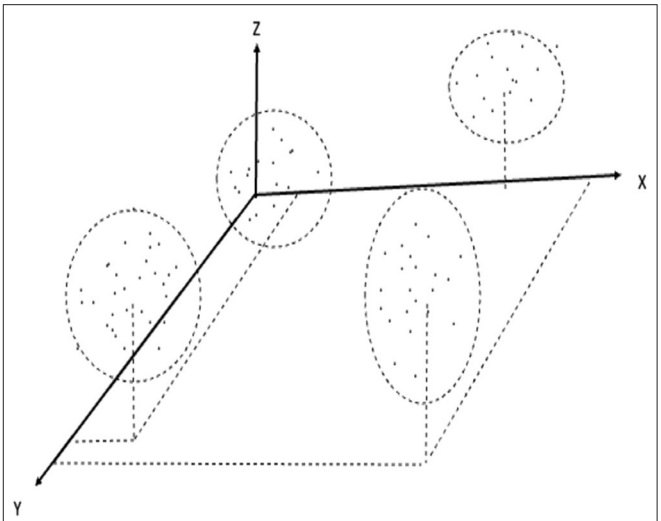

FIGURE 12 | Example of clustering.

Hersey and Blanchard suggest four different leadership styles.

- If the learner is low in competence and high in commitment (development level D1) the theory suggests *Directing* feedback, i.e., high directive and low supportive.
- If the learner has some competence but low commitment (development level D2) the theory suggests *Coaching* feedback, i.e., high directive and high supportive.
- If the learner has moderate to high competence but lacking commitment (development level D3) the theory suggests *Supporting* feedback, i.e., low directive and high supportive.
- If the learner has a high degree of competence and a high degree of commitment (development level D4) the theory suggests *Delegating* feedback, i.e., low directive and low supportive.

A popular concept in the behavioral sciences with respect to the form of provided feedback is *Nudging* (Thaler and Sunstein, 2008). Nudging is seen as a technique that suggests positive reinforcements and indirect suggestions as ways to create favorable behavior and good decision making. This form seems to be in accordance with *Supporting feedback* as defined by Hersey and Blanchard.

Another perspective on learning in a rich extended learning environment including an AI system is through *Technopedagogy*. Technopedagogy is the pedagogical considerations uniquely associated with the integration of digital technology (Newson, 1999). Emphasis is on tailoring technology to suit pedagogy, rather than tailoring pedagogy to suit technology. Such digital technology should also foster connections and facilitate for the participants in the learning environment to connect with each other. In such an environment, it should be easy for all agents to engage and disengage with technology when appropriate. Cause even though digital technology has the power to connect, digital technology also has the power to distract.

IRT—Item Response Theory

Most of the feedback will be to the learner. Feedback to the other agents in the learning system may not require that much focus. Much of the feedback created will nevertheless be presented both to the learner and some of the other agents (maybe simultaneously), as it also may be informative to them. In for example Item Response Theory (IRT) both the ability level of the learner and a characterization of the different questions in an assessment are estimated, which constitute information important both to the learner and to a teacher.

The objective of item response theory (IRT) is to characterize test items and estimate the ability of an examinee (Embretson and Reise, 2000). The basic idea is to estimate the probability that an examinee provides a correct response to items presented in a questionnaire. This probability of correct response is assumed to be a function of an underlying trait or ability, θ. θ is modeled as a stochastic variable typically depicted as ranging from -3 to 3. Usually, the probability distribution of θ is assumed to be

$$\theta \sim \mathrm{N}(\mu, \mathrm{s}^2 = 1^2).$$

An estimate of θ to the left of the expectation μ in this probability distribution reflects that the learner is in the lower half of the population with respect to ability. An estimate of θ equal to the expectation μ reflects that the learner is "an average" learner in the population, while an estimate of θ larger than μ reflects that the learner is in the upper half of the population with respect to ability.

The Item Response Function (IRF) gives the probability that a learner j with a given ability level θ_j will answer correctly on item *i*. As θ increases, the probability of a correct response $p_i(\theta_j)$ increases as modeled in the following function.

$$p_i(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

where

- $a =$ Discrimination index ("slope").
- $b =$ Difficulty index.
- $c =$ Lower asymptote ("guessing").

Presenting feedback to the learning system based on IRT may be as illustrated in **Figures 13**, **14**.

In **Figure 13** the dashed curve is representing an "easy" item ($b = -1$) and the dotted curve a "difficult" item ($b = 1$). When an item is represented by the dashed curve the probability of a correct answer to this item is $\approx$0.80 for a learner with an ability corresponding to a value $\theta = 0$. For an item represented by the dashed curve the probability of a correct answer to this item is $\approx$0.44 for a learner with an ability corresponding to a value $\theta = 0$. Thus, this item is estimated to be more difficult.

In **Figure 14** we can see that we expect "no one" with an ability level slightly below 0 to get the item represented by the solid line ("large" $a = 6$) correct, while we at the same time expect "everybody" with an ability level slightly better than 0 to get it correct. (*If you're below average, you won't make it. If you're above average, you're quite certain to make it.*) That is, this question is probably discriminating too much, which may suggest the teacher to revise the question.

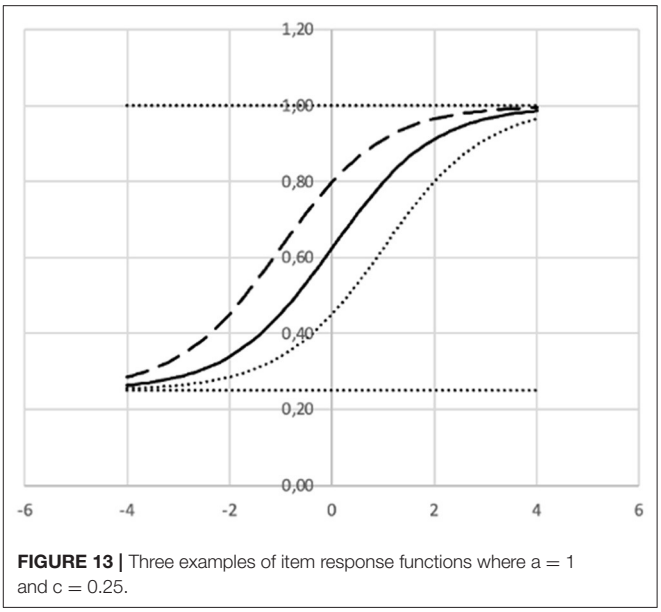

FIGURE 13 | Three examples of item response functions where a = 1 and c = 0.25.

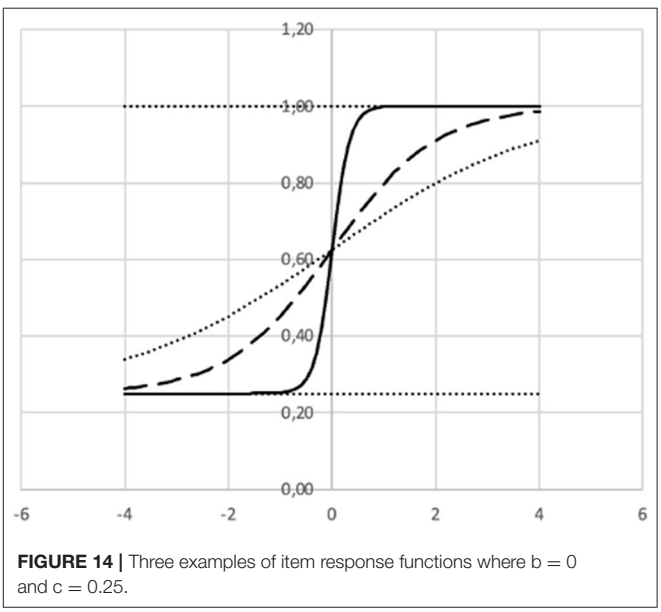

FIGURE 14 | Three examples of item response functions where b = 0 and c = 0.25.

The item represented by the dotted line is kind of "easier" for the "not so smart," but still difficult for the "smart ones" ("small" $a = 0.5$). That is, the item is not very discriminating.

Meta Learning

Meta learning was originally introduced by Maudsley (1979) and later used by Biggs (1985) to describe the state of being aware of and taking control of one's own learning. A learner needs a sufficient high level of situational awareness to be able to assess the effectiveness of her own learning approach and modify it according to the demands of the learning task. Meta

learning, being an active, internal process, also relates to learners' attitudes, such as their belief that the way they adapt to the learning situation is the best way for them, and that they have the capacities and confidence to apply their knowledge. Meta learning can also be an effective tool in assisting students to become independently self-reflective (Biggs, 1985).

CONCLUSIONS

During the theoretical considerations while development an XAI system to improve the learning process by adapting to the individual learner, some lessons are learned.

It seems fundamental to see the objective of a learning process to be to increase both the learner's

- competence,
- confidence,
- learning ability.

This threefold objective is both important with respect to the evaluation of a possible improvement of the learning process, as it is suggesting that there may be new forms of feedback to the learning system in addition to those directly connected to improving competence.

Realizing that the complete learning environment should include more than a learner and a teacher is also key for success. All the resources within the rich extended learning environment should be utilized both for establishing and maintaining distributed situational awareness and to improve the distributed cognition in this system to accomplish the objectives.

With these lessons learned it should be possible to introduce an AI system into a learning process and improve the learning process by adapting to the individual. It should be possible to both

- improve the learning process for many learners,
- make adaptive learning more easily accessible,
- empower teachers,
- improve education management and delivery,
- offering life-long opportunities for all, making the delivery of education more democratic.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This research was funded in part by The Research Council of Norway [310123]. A CC BY or equivalent license is applied to any Author Accepted Manuscript (AAM) version arising from this submission, in accordance with the grant's open access conditions.

REFERENCES

Aarset, M. (2014). *Risk, Issues and Crisis Management*, ebook, terp.no. Haugesund, Norway: TERP AS.

Aarset, M. V., and Glomseth, R. (2019). "Police leadership during challenging times," in *Policing and Minority Groups*, eds J. F. Albrecht, G. den Heyer, and P. Stanislas (Cham: Springer), 29–53.

Anderson, L. W., and Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities, and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Barley, S. R. (2020). *Work and Technological Change*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198795209.001.0001

Biggs, J. B. (1985). The role of meta-learning in study process. *Br. J. Educ. Psychol.* 55, 185–212. doi: 10.1111/j.2044-8279.1985.tb02625.x

Bloom, B. S. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York, NY: David McKay Company.

Corno, L. (1993). The best-laid plans. *Educ. Res.* 22, 14–22. doi: 10.2307/1176169

Corno, L., and Kanfer, R. (1993). The role of volition in learning and performance. *Rev. Res. Educ.* 19, 301–341. doi: 10.3102/0091732X019001301

Cox, D. R., and Miller, H. D. (1987). *The Theory of Stochastic Processes*. London: Chapman and Hall.

Dede, C. J., Jacobson, J., and Richards, J. (2017). "Introduction: virtual, augmented, and mixed realities in education," in *Virtual, Augmented, and Mixed Realities in Education*, eds D. Liu, C. Dede, R. Huang, and J. Richards (Singapore: Springer), 1–16.

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. New York, NY: Psychology Press.

Garcia, T., McCann, E. J., Turner, J. E., and Roska, L. (1998). Modeling the mediating role of volition in the learning process. *Contemp. Educ. Psychol.* 23, 392–418. doi: 10.1006/ceps.1998.0982

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York, NY: The Guilford Press.

Hersey, P., and Blanchard, K. (1969). Life cycle theory of leadership. *Train. Dev. J.* 23, 26–35.

Herzberg, F. I. (1982). *The Managerial Choice: To Be Efficient and to Be Human, 2nd Edn*. Salt Lake City, UT: Olympus Publishing Company.

Hollnagel, E., and Woods, D. D. (2005). *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. Boca Raton, FL: CRC Press. doi: 10.1201/9781420038194

Hutchins, E. (2001). "Cognition, distributed," in *International Encyclopedia of the Social and Behavioral Sciences*, eds N. J. Smelser and P. B. Baltes (Amsterdam: Elsevier), 2068–2072.

Illeris, K. (2009). *How We Learn: Learning and Non-learning in School and Beyond*. New York, NY: Routledge.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: Wiley.

Kolb, D. A. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Hoboken, NJ: Prentice Hall.

Marca, D. A., and McGowan, C. L. (1988). *SADT: Structured Analysis and Design Technique*. New York, NY: McGraw-Hill.

Maudsley, D. B. (1979). *A Theory of Meta-learning and Principles of Facilitation: An Organismic Perspective*, *Vol. 40*. Toronto, ON: University of Toronto, 4354–4355-A.

McCrae, R. R. (2018). "Defining traits," in *The SAGE Handbook of Personality and Individual Differences*, eds. V. Zeigler-Hill, and T. K. Shackelford (London: SAGE), 3–22.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press.

Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology*. San Francisco, CA: Freeman.

Newson, J. (1999). Techno-pedagogy and disappear in context. *Academe* 85, 52–55. doi: 10.2307/40251770

Norman, D. A. (1991). "Cognitive Artifacts," in *Cambridge Series on Human-Computer Interaction, No. 4. Designing Interaction: Psychology at the Human-Computer Interface*, ed J. M. Carroll (New York: Cambridge University Press), 17–38.

Piaget, J. (1973). *Psychology and Epistemology: Towards a Theory of Knowledge*. Rio de Janeiro: Record.

Salmon, P. M., Stanton, N. A., Walker, G. H., and Jenkins, D. P. (2009). *Distributed Situation Awareness: Theory, Measurement and Application to Teamwork*. Surrey: Ashgate.

Salomon, G. (1997). *Distributed Cognition: Psychological and Educational Considerations*. New York, NY: Cambridge University Press.

Sternberg, R. J. (1994). "Thinking styles: Theory and assessment at the interface between intelligence and personality," in *Personality and Intelligence*, eds R. J. Sternberg, and P. Ruzgis (New York, NY: Cambridge University Press), 105–127.

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Thompson, G., and Aarset, M. (2012). Examining the impact of social intelligence, demographics, and context for implementing the dynamics of the situational leadership model. *J. Int. Doct. Res.* 1, 122–142. Available online at: http://hdl.handle.net/11250/93920

Um, E. "R.," Plass, J. L., Hayward, E. O., and Homer, B. D. (2012). Emotional design in multimedia learning. *J. Educ. Psychol.* 104, 485–498. doi: 10.1037/a0026609

Vygotsky, L. S. (1978). *Mind in Society*. Cambridge, MA: Harvard University Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 22 July 2022
DOI 10.3389/frai.2022.941825

OPEN ACCESS

EDITED BY
Paul Seitlinger,
University of Vienna, Austria

REVIEWED BY
Gerti Pishtari,
Danube University Krems, Austria
Helena Macedo Reis,
Federal University of Paraná, Brazil

*CORRESPONDENCE
Gautam Biswas
gautam.biswas@vanderbilt.edu

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 11 May 2022
ACCEPTED 27 June 2022
PUBLISHED 22 July 2022

CITATION
Vatral C, Biswas G, Cohn C, Davalos E and Mohammed N (2022) Using the DiCoT framework for integrated multimodal analysis in mixed-reality training environments.
Front. Artif. Intell. 5:941825.
doi: 10.3389/frai.2022.941825

Using the DiCoT framework for integrated multimodal analysis in mixed-reality training environments

Caleb Vatral, Gautam Biswas*, Clayton Cohn, Eduardo Davalos and Naveeduddin Mohammed

Open Ended Learning Environments, Department of Computer Science, Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN, United States

Simulation-based training (SBT) programs are commonly employed by organizations to train individuals and teams for effective workplace cognitive and psychomotor skills in a broad range of applications. Distributed cognition has become a popular cognitive framework for the design and evaluation of these SBT environments, with structured methodologies such as *Distributed Cognition for Teamwork (DiCoT)* used for analysis. However, the analysis and evaluations generated by such distributed cognition frameworks require extensive domain-knowledge and manual coding and interpretation, and the analysis is primarily qualitative. In this work, we propose and develop the application of multimodal learning analysis techniques to SBT scenarios. Using these analysis methods, we can use the rich multimodal data collected in SBT environments to generate more automated interpretations of trainee performance that supplement and extend traditional DiCoT analysis. To demonstrate the use of these methods, we present a case study of nurses training in a mixed-reality manikin-based (MRMB) training environment. We show how the combined analysis of the video, speech, and eye-tracking data collected as the nurses train in the MRMB environment supports and enhances traditional qualitative DiCoT analysis. By applying such quantitative data-driven analysis methods, we can better analyze trainee activities online in SBT and MRMB environments. With continued development, these analysis methods could be used to provide targeted feedback to learners, a detailed review of training performance to the instructors, and data-driven evidence for improving the environment to simulation designers.

KEYWORDS

distributed cognition, learning analytics (LA), multimodal data, simulation based training (SBT), mixed reality (MR), DiCoT, human performance, multimodal learning analytics (MMLA)

1. Introduction

Modern workplaces require workers to develop and execute a complex combination of cognitive, metacognitive, and psychomotor skills to achieve effective performance. With advanced technologies that have now become widely available, faster and more effective skill development can be achieved by designing effective training protocols that provide learners with multiple opportunities to train along with formative feedback to support continual improvement with clear pathways to achieve proficiency in their tasks. Simulation-based training (SBT) has become a popular paradigm to implement these training protocols. These environments provide safe and repeatable spaces for learners to practice and develop their workplace skills, and combined with adequate debrief and feedback they can support training in multiple domains (Ravert, 2002; Gegenfurtner et al., 2014).

When SBT scenarios require collaboration and feedback among multiple agents (real and virtual), it is common to interpret the training scenarios and trainee performance using theories of *distributed cognition* (Hollan et al., 2000; Hutchins, 2000). Furthermore, many SBT environments incorporate physical movement and embodiment, teamwork behaviors, and domain-specific tools to aid the workers, which match with the core tenets of distributed cognition (Kaplan et al., 2021). This is especially the case for SBT environments that are enhanced using mixed-reality tools, in domains such as emergency response, collaborative and embodied learning, and healthcare (Rosen et al., 2008; Mirchi et al., 2020; Rokhsaritalemi et al., 2020). Techniques such as *Distributed Cognition for Teamwork (DiCoT)* have been successfully applied to analyze SBT, both for the purposes of simulation design and learner feedback (Hazlehurst et al., 2008; Rybing et al., 2016, 2017). Traditionally, analysis of distributed cognition with these frameworks relies heavily on human observations by researchers and domain experts to provide a descriptive analysis of performance in the learning and training scenarios.

In parallel, other learning domains, such as K-12 classrooms, have seen a transformation in personalized learning through data-driven learner modeling and multimodal learning analytics (Hoppe, 2017; Ochoa et al., 2017). In these learning environments, data from student interactions are logged and analyzed to produce insights into the learners' cognitive, metacognitive, and affective processes, and the impact these processes have on their learning outcomes. While learning analytics has been employed to analyze learner performance in some simulation-based training domains as well, for example, in Biswas et al. (2019), Kim et al. (2018), and Martinez-Maldonado et al. (2020a), these applications are less common and often rely on cognitive theories derived from traditional learning frameworks. For learning and training in mixed reality-based simulation environments that involve multiple agents and combination of physical and virtual spaces, more advanced cognitive theories, such as distributed cognition, better match the affordances provided by the environments.

Motivated by this gap, in this paper we develop a framework to apply a *mixed quantitative + qualitative* approach that combines multimodal data analysis in the context of distributed cognition to analyze learner behavior and performance in SBT environments. In particular, our studies focus on a mixed-reality manikin-based (MRMB) environment for training nurses to work with patients in hospital rooms. MRMB-based simulation training provides realistic and high-fidelity scenarios for nurses to train in. They have proven to be quite effective in helping nurses develop and achieve proficiency in psychomotor, cognitive, and social skills as they interact with patients and equipment, make diagnoses, and provide interventions to alleviate their patient's problems (Hegland et al., 2017).

As a demonstration of our framework for tracking and analyzing trainee behaviors and performance, we ran a small study with nursing students in this MRMB training environment. We have developed and applied our mixed quantitative + qualitative methods approach to analyze the data collected with video, audio, and eye tracking sensors. Our computational architecture processes the raw multimodal data streams and analyzes this data framed using the constraints and insights derived from a qualitative analysis using the DiCoT distributed cognition approach. The results are mapped to a combined qualitative-quantitative representation of the nurses' problem solving behaviors and performance, with the help of our cognitive task model. With continued development and refinement, results from our analysis methods can be provided to learners as formative feedback and to instructors to help them guide more detailed discussions during simulation debriefing.

The analysis presented in this paper supports an investigation of two primary research questions:

1. How can multimodal learning analysis be used to support a comprehensive analysis of distributed cognition in MRMB simulation training environments?
2. How does temporal alignment and analysis of multiple data modalities help us gain a deeper understanding of trainees' actions in the context of the tasks they are performing in an MRMB environment?

The rest of this paper is organized as follows. Section 2 presents previous work on SBT, the Distributed Cognition framework, and an overview of multimodal data analysis approaches applied to studying learner behaviors. Section 3 discusses our theoretical framing of the training scenarios and analysis by combining cognitive task modeling, distributed cognition through the DiCoT methodology, and multimodal data analytics. Section 4 provides details of the methods we have adopted in our study; first an overview of the MRMB-based Nurse Training scenario, a Cognitive Task Analysis approach

to interpreting and analyzing nurses' actions in the training environment and mapping them to higher level cognitive behaviors, our adaptation of the DiCoT framework to study nurse performance and behaviors in the training scenarios, and a complete computational architecture to derive performance analysis from data collected in the SBT environment. Section 5 presents details of the analyzes of the nurses' performance and behaviors in the case-study MRMB-based training environment. This is followed by a discussion of the results obtained for two the scenarios and their broader implications in Section 6. Last, Section 7 provides the conclusions of the paper, limitations with the current approach, and directions for future work.

2. Background and related work

In this section, we briefly review past work in SBT, distributed cognition, and multimodal analytics applied to analyzing learners' training performance and behaviors.

2.1. Simulation-based training

Simulation-based environments offer many attractive properties for training applications; they provide controllable and repeatable environments in which learners and trainees can safely practice complex cognitive and psychomotor skills in rich and dynamic scenario representations. Thus, it is not surprising that simulation-based training has been widely adopted for a variety of domains, and many studies have shown them to be effective for both training and assessment (Ravert, 2002; Maran and Glavin, 2003; Daniels and Auguste, 2013; Rybing, 2018). In medical domains, SBT has been used since the 1950s when the first commercial medical training manikin was released. The manikin-based approach combined with computer-based simulations continues to be widely utilized and studied today (Cooper and Taqueti, 2008; Hazlehurst et al., 2008; Pimmer et al., 2013; Rybing et al., 2017). For example, Rybing et al. (2017) studied the use of simulation-based training for nurses in mass causality events; Kunst et al. (2016) studied the use of manikin simulation for mental health nursing; and Johnson et al. (2014) found that mankin-based education was more effective than web-based education for advanced practice nursing students. For further information, see Cooper and Taqueti (2008) which reviewed the history and development of manikin-based clinical education, Al-Ghareeb and Cooper (2016) which reviews the current state of manikin-based clinical education along with its barriers and enablers, and Gegenfurtner et al. (2014) which reviewed the larger context of digital simulation-based training.

In addition, the integration of simulation environments with advanced computing resources has led to further advances in the field. Computer-based simulations allow for automated collection of trainee activity data, which can then be used to evaluate their performance, and for debriefing and after-action reviews (Ravert, 2002; Sawyer and Deering, 2013). In medical domains, a lot of the computer-based simulation training relies on high fidelity manikins that trainees can realistically interact with to practice their clinical and teamwork skills (Al-Ghareeb and Cooper, 2016). This creates *mixed-reality* environments, where trainees act in a physical space, which includes real equipment that interfaces with a digital simulation. The digital simulation controls the patient manikin's vital signs and overall health manifestations. In addition, the digital simulation can take into account trainees actions in the environment and on the manikin, and adapt the manikin's vital signs and responses to these actions.

The overall goal of SBT is to help learners to develop a set of skills that are *transferable*, meaning the skills acquired in the simulation can be utilized in other simulation settings and in real-world situations. In particular, one of the primary goals for medical SBT is to help trainees develop skills that transfer from the simulation environments to actual medical settings with real patients. *Application validity* measures capture how well SBT environments accomplish this transfer for a sufficiently large population of trainees (Feinstein and Cannon, 2002).

Prior work has shown that providing formative feedback during debrief after the simulation improves both the application validity of the simulation, as well as the competence and self-efficacy of the learners (Gegenfurtner et al., 2014). It is important to note that the formative feedback provided must be discussion and explanation focused, and not purely evaluative in order to preserve the psychological safety of the training environment (Kang and Min, 2019). While similar simulation environments are also used for learner assessment (Cook et al., 2014), our focus in this paper is on simulation-based *training*, where learners must feel safe to practice and not fear that mistakes will have long-term negative consequences (Kang and Min, 2019; Park and Kim, 2021). Taking this into account, our work focuses on building analysis methods designed to provide feedback that will guide and support discussion and learning during debrief. Our analysis methods are based on multimodal data generated by the mixed-reality environment grounded in the theory and practice of distributed cognition.

2.2. Distributed cognition

Traditionally, cognition is studied with the individual as the basic unit of analysis. In essence, this classical view of cognition views the brain of an individual as a processing unit, which takes input from the outside world, manipulates this information, and produces some output, often in the form of body functions, such as movement and speech (Clark, 1997). However, this view of an individual mind as the basic unit of cognition ignores the complex relationship between the mind, the body, and the larger environment. The ability

to leverage movement, tools, technology, collective wisdom, and social structures allows humans to achieve far more than an isolated individual mind alone can, but the traditional view of cognition marginalizes these embodied, cultural, and environmental components (Geertz, 1973; Hazlehurst et al., 2008).

These limitations with classical cognition led some cognitive scientists, such as Clark, Hutchins, Cole, and others in the late twentieth century to begin developing alternative systems of examining cognition (Hutchins, 1995; Clark, 1997; Cole, 1998). One such alternative approach is *Distributed Cognition*, developed by Hutchins and colleagues (Hutchins, 1991, 1995, 2000, 2006). Distributed cognition (DCog) extends the boundaries of classical cognition from the mind of an individual in isolation into a collective that includes the individual's mind, body, other people, and the environment in which the cognition is taking place. Instead of the unit of cognitive analysis being the individual mind, distributed cognition treats the entire activity system as the unit of analysis, with the goal of understanding cognition at this system level (Hazlehurst et al., 2008; Rybing, 2018).

Hutchins argues that cognition occurs across at least three different modalities (Hutchins, 2000). First, cognition can be distributed across members of a *social group*. This can be seen as individuals coming together to solve a problem and contribute to a common goal. Second, cognition can be distributed between *internal* and *external structures*. This is most evident in the use of tools, where individuals offload some cognitive processing to a material or environmental object, but can also have some less apparent manifestations, such as the layout of a physical space affecting cognition. Third, cognition can be distributed across *time*, with the nature and outcomes of earlier events affecting the nature of later events (Hutchins, 2000).

Distributed cognition is particularly relevant in analyzing training performance and behaviors in mixed-reality, simulation-based training. Mixed-reality SBT environments manifest many of the characteristics of these three distributed modalities. SBT inherently contains social structures and roles over which cognition is distributed. When multiple learners train simultaneously in the environment, the social distribution and interactions can be studied explicitly, with the learners collaborating and sharing the cognitive load and decision making processes in the task. Even in SBT cases with only one learner, there is a social distribution between the learner and the instructor, with information traveling and transforming between the instructor and student as they interact. SBT also contains instances of cognition distributed between internal and external structures. In mixed-reality scenarios, there is a distribution between the learners' minds, the physical space they inhabit, and the digital space with accompanying interfaces that are controlled by the simulation. In addition, many training domains require learners to learn and operate domain-specific tools, which also represent artifacts of distributed cognition. Finally, SBT is necessarily temporal, as learners practice skills that change (improve or degrade) over time. Thus, previous practice and previous actions will affect the ways in which learners approach current cognitive tasks.

Other studies which focus on nursing simulation-based training have also adopted distributed cognition for their analysis. Rybing et al. (2017) use distributed cognition to analyze nursing students training on a mass causality simulation; Pimmer et al. (2013) contrast various cognitive theories used in clinical learning to highlight advantages of distributed cognition; and (Hazlehurst et al., 2008) discuss the use of distributed cognition as a framework for medical informatics. Because of this overlap between the distributed cognition framework and the modeling and interpretation of learner behavior in simulation based training in general, and in medical and nurse training in particular, we ground our analysis methods using Distributed Cognition as a theoretical framework.

2.3. The DiCoT analysis framework

Despite the advantages of distributed cognition as a cognitive framework, application of the framework requires specific methodologies that are not outlined in the original work. Several structured qualitative analysis methodologies have been developed for analyzing distributed cognition in different domains and scenarios. For example, Wright et al. (2000) proposed the *Resource Model* to study human computer interaction in a team framework, Galliers et al. (2007) proposed the *Determining Information Flow Breakdown (DIB)* model to study organizational learning in response to adverse events in medical settings, and Stanton (2014) proposed the *Event Analysis of Systemic Teamwork (EAST)* framework that employs three network models (i.e., task, social and information) to analyze the interactions between the sound room and control room in a submarine. Following the wide adoption of distributed cognition models and their success in analyzing trainee behaviors in the medical training domain (e.g., Hazlehurst et al., 2008; Pimmer et al., 2013; Rybing et al., 2017), we adopt the *Distributed Cognition for Teamwork (DiCoT)* model proposed by Blandford and Furniss (2006).

DiCoT is a qualitative analysis framework designed to analyze distributed cognition by breaking down a cognitive system into five independent themes: (1) *information flow*, (2) *artifact and environment*, (3) *physical layout*, (4) *social interactions*, and (5) *temporal evolution* (Blandford and Furniss, 2006). The information flow model focuses on how information propagates and transforms through the system. The artifact theme follows how tools can be used to aid the cognition of the system. The physical layout theme examines the way in which objects and people are arranged in a space and how that arrangement affects cognition. The social model focuses on the relationships between people in the cognitive

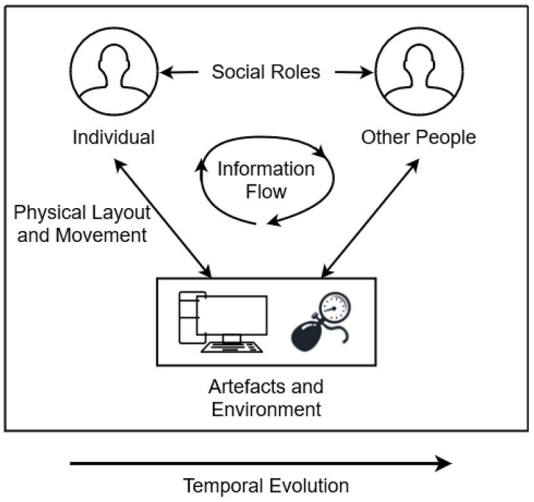

FIGURE 1
Illustration of the interactions between each of the five DiCoT themes and how they work together to construct the entire cognitive system.

system and the individual's differing knowledge, skills, and abilities. Finally, the temporal evolution model focuses on how the system changes over time. Each of the five DiCoT themes contributed components to our understanding the overall cognitive processes and psychomotor skills that trainees employed in the environment, but the themes are also highly interconnected. Figure 1 illustrates how the themes manifest within a cognitive system and the various ways in which the different themes interact with one another. For example, social roles mediate how information flows between different individuals; physical layout mediates how information flows between individuals and the environment; and temporal evolution describes and mediates how these processes change over time. By analyzing each of the themes individually and how each theme interacts with the others, we can fully understand the distributed cognition processes and psychomotor skills being enacted in the system.

In order to analyze each of the themes and their interactions, the DiCoT methodology defines several *principles* that describe the ways in which each component of the system contributes to the overall cognitive process. For example, principle 10: Information Hubs, describes that certain artifacts in the system are central focuses where different channels of information meet. This principle is primarily related to the *information flow* and *artifacts and environment* themes. By analyzing the different artifacts in a distributed cognitive system and how they are referenced for information, we can determine which artifacts represent information hubs and how the design of those hubs influences the overall cognition of the system. Each of the 18 principles is analyzed in a similar way, but relate to other components of the system. All eighteen DiCoT principles are summarized in Table 1. By analyzing the distributed cognition system and identifying the manifestations of each of these 18 principles within the system, we can understand how each of the 5 DiCoT themes work together to construct the overall cognition of the system. We discuss our qualitative analysis of the nurse training simulation using DiCoT framework in Section 4.3.

2.4. Multimodal learning analysis

Learner modeling based on student performance and behavior has been the cornerstone for adapting and personalizing computer-based learning environments to individual learner needs. More recently, data-driven approaches to learner modeling based on learning analytics and machine learning methods have become popular for capturing and analyzing learner behaviors in complex instructional and training domains (Hoppe, 2017). With the development of these data-driven learning analytics techniques gives rise to the question: *What forms of data need to be collected to enable meaningful analysis in specific learning scenarios*? In traditional computer based learning environments, typical data collection includes interactions with the system that can be logged. Analysis of the logged data paints a reasonable picture of the learners' activities in the context of the tasks they are performing in the environment (Hoppe, 2017; Ochoa et al., 2017).

However, more recent work has begun to point out the potential limitations of these traditional methods. By only using logged data that is easy to collect, we may miss out on important context and interpretation that the information sources may provide. Therefore, we may require additional sensors to collect such data (Ochoa et al., 2017). In other words, to aviod the so-called *streetlight effect* (Freedman, 2010), researchers have begun to consider alternative and more complex data sources, such as physical movement, gestures, and posture captured with video; dialogue captured using microphones; stress levels captured with biometric sensors; and gaze and attention collected using eye tracking devices. Data collected using these modalities become especially important when the learning or training task requires operations in physical or mixed-reality spaces, and when learners work in groups to accomplish overall goals.

Combining all of the modalities of operation (e.g., activities, communication, affective states, stress levels, and gaze) can lead to analyzes that provide a more complete picture of the cognitive, psychomotor, and metacognitive processes of the learners (Blikstein and Worsley, 2016). The focus on collection, processing, and analysis of this quantity and variety of data has been the basis for new research and analyzes in the field of multimodal learning analytics (MMLA) (Blikstein, 2013; Blikstein and Worsley, 2016; Worsley and Martinez-Maldonado, 2018). These new MMLA methods have also been applied to simulation-based training environments. For

TABLE 1 The 18 principles of DiCoT analysis, summarized from Blandford and Furniss (2006).

Principle name	Description
1. Space and cognition	The role space and spatial layout play in supporting cognition
2. Perceptual principle	Spatial representations support cognition more than non-spatial representations, as long as there is a clear mapping between the space and that which the space represents
3. Naturalness principle	Cognition is aided when the form of a representation matches the properties of what it represents
4. Subtle bodily supports	Individuals often use their body to support cognitive processes
5. Situation awareness	People need to be informed of and understand what has previously happened, what is currently going on, and what is planned
6. Horizon of observation	The information that can be seen or heard by a person; closely related to and influencing situation awareness
7. Arrangement of equipment	The layout of equipment affects what information people have access to, and thus their ability to process it
8. Information movement	Information moves around a system in a number of ways, which all have unique functional consequences
9. Information transformation	Information can be represented in many forms, and often must transform between these forms when moving and when being processed
10. Information hubs	A central focus or source where different channels of information meet and are processed together
11. Buffering	If incoming information interferes with ongoing activities, buffering allows the information to be held until an appropriate time where it will not interfere
12. Communication bandwidth	Different modalities of communication often carry different amounts of information. For example, face-to-face communication offers more information than computer-mediated communication
13. Informal communication	Not all communication is formal, and sometimes informal communication can carry very important information that is not otherwise passed
14. Behavioral trigger factors	Groups of people can operate together without an overall plan by individually responding appropriately to certain local trigger factors
15. Mediating artifacts	People often bring artifacts into coordination to support completion of a task
16. Creating scaffolding	People often simplify their cognitive tasks by utilizing their environment
17. Representation-Goal Parity	When an artifact is used to represent the system's goal, representations closer to the goal of the user are more powerful
18. Coordination of Resources	Different information structures can be coordinated to aid in cognition

example, Martinez-Maldonado et al. (2020b) examined how to design actionable learning analytics for manikin-based nurse training; Di Mitri et al. (2019) designed MMLA methods for detecting mistakes during CPR training; and López et al. (2021) studied collaborative behaviors in serious tabletop games using MMLA methods.

In our own previous work, we have applied MMLA methods to analyze teamwork behaviors in simulation-based training environments, including those that incorporate mixed-reality components (Biswas et al., 2019; Vatral et al., 2021, 2022). Our analyzes of learner performance and behaviors have been based on *cognitive task analysis*, which is a set of methods commonly used to describe and decompose complex problem-solving domains into their core cognitive proficiencies (Clark and Estes, 1996; Schraagen et al., 2000; Zachary et al., 2000). These cognitive components describe multiple parameters that include goal setting, planning, decision making, declarative and procedure knowledge and execution, and situational awareness (Militello and Hutton, 1998). The models and insights generated from the task analysis are often critical in the design and development of training systems for these complex domains.

2.5. Cognitive task analysis

Cognitive Task Analysis typically draws from multiple sources. This includes a review of relevant literature, interviews with domain experts, and observing and interpreting training activities in the mixed reality simulation environment in terms of their conceptual and procedural components. From this analysis, one can build a comprehensive task-subtask hierarchy that links high-level tasks and subtasks down to specific observable skills and activities performed by trainees (Biswas et al., 2019; Vatral et al., 2021). The hierarchy is designed to support the inference of complex cognitive concepts by analyzing observable behaviors and data. Cognitive processes related to task execution are located at the highest level of the task hierarchy, and each deepening level representing more

concrete and observable manifestations of these concepts within the task domain.

By analyzing the observable multimodal data at the lowest levels of the hierarchy and propagating the results up to higher levels, we can generate inferences about cognitive activities and competencies of trainees. In this way, insights generated from cognitive task analysis combine top-down model-driven and bottom-up data-driven approaches. In previous work, we have applied cognitive task analysis methods to demonstrate how teamwork in mixed-reality SBTs can be evaluated using MMLA (Vatral et al., 2022). In this paper, we extend this work and further ground the MMLA analyzes methods in distributed cognition, as described in the next section.

3. Theoretical framework

Our goal in this work is to present a framework for combining the benefits and insights from qualitative analysis of distributed cognition through the DiCoT methodology and quantitative analysis through data-driven multimodal analytics. Analysis using qualitative methods (Cognitive Task Analysis, DiCoT) provides domain semantics to inform how the quantitative analysis (MMLA) is performed, and in turn, results of the quantitative analysis provide new insights into the domain and the learner behaviors that inform a richer qualitative analysis. We believe that by presenting an integrated qualitative and quantitative analysis that inform and shape one another, the strengths of each method can be amplified, thus providing for a deeper insights than each method individually and better feedback to learners, instructors, simulation designers, and researchers.

Our overall theoretical framework, illustrated in Figure 2, breaks down this cyclic combined qualitative and quantitative analysis approach into three major components: DiCoT analysis, multimodal analytics, and the cognitive task model. The cognitive task model provides the cornerstone of the overall framework. For our MRMB simulation environment, we frame the task model around the set of primary tasks that define the training or learning domain. These concepts represent the mapping of the task domain into the overarching cognitive processes, psychomotor skills, affective states, and collaborative processes that are relevant to the task domain. For example, learning and training domains typically include high level cognitive processes such as information acquisition, problem solving, solution construction, solution testing, and evaluation. These processes, in a broad sense, remain invariant across multiple domains and training scenarios. However, their interpretation and execution may differ depending on the training scenario and the domain under consideration.

Next, we construct the hierarchical structure by breaking down the highest level cognitive, psychomotor, affective, and collaboration concepts into their more domain specific sub-components and sub-tasks using a progressive elaboration process. The primary reason for creating the different levels of abstraction is to ensure that variations of training scenarios, though they may differ in their lower-level task definitions and sub-divisions, map onto relevant higher level processes and help define proficiency measures in the task domain.

In more detail, primary tasks are decomposed into sub-tasks; sub-tasks are further decomposed into more fine-grained sub-tasks; and so on until we reach a set of basic task units that cannot be meaningfully decomposed further. We call this basic unit an *action*. Each sub-task represents a constituent requirement that is sequenced and completed to accomplish the higher-level tasks in the layer above them. Moving toward the lower levels of the hierarchy, the sub-tasks become more and more domain-specific, and at the lowest levels map onto observable actions and behaviors. For example, consider information-acquisition as the highest-level cognitive task. In order to acquire information, we may visit a library, search the internet, ask a friend, and so on. The specific sub-tasks included within the task model are limited by the domain being analyzed. By limiting each level to sub-tasks specific to the given domain, we follow a top-down approach to modeling and produce a task space model of the domain.

While the modeling of the domain is approached top-down, the interpretation of the learner actions and behaviors uses the model in a bottom-up manner, interpreting the multimodal data collected from the environment into lower level activities and behaviors. We employ a variety of multimodal analysis techniques to link from observable data back to the interpreted actions performed by the learners. This is illustrated by the arrow linking multimodal analytics (green) to the cognitive task model (blue) in Figure 2. The specific analytics and algorithmic methods utilized depend on the domain being analyzed and the specific sensors that are available. For example, if microphones are available, we can apply natural language processing algorithms to convert the audio signals to semantic information on the topic of the conversation. Similarly, if we collect video data, then computer vision techniques can be used to understand movement actions within the simulation space. The design of these analytics and algorithmic methods within a specific domain are informed by the qualitative DiCoT analysis, as illustrated by the arrow linking distributed cognition (yellow) to multimodal analytics (green) in Figure 2. By analyzing the training environment using the DiCoT methodology, we can determine important components of the task domain that inform the categories and classes for the algorithmic models.

For example, in our nursing domain, the DiCoT analysis revealed that there are four meaningful semantic areas in the simulation space: left of the bed, right of the bed, foot of the bed, and outside the room (see Section 4.3.1). Thus, we can adopt this result from the qualitative analysis into the design of the quantitative algorithmic methods by using the video data

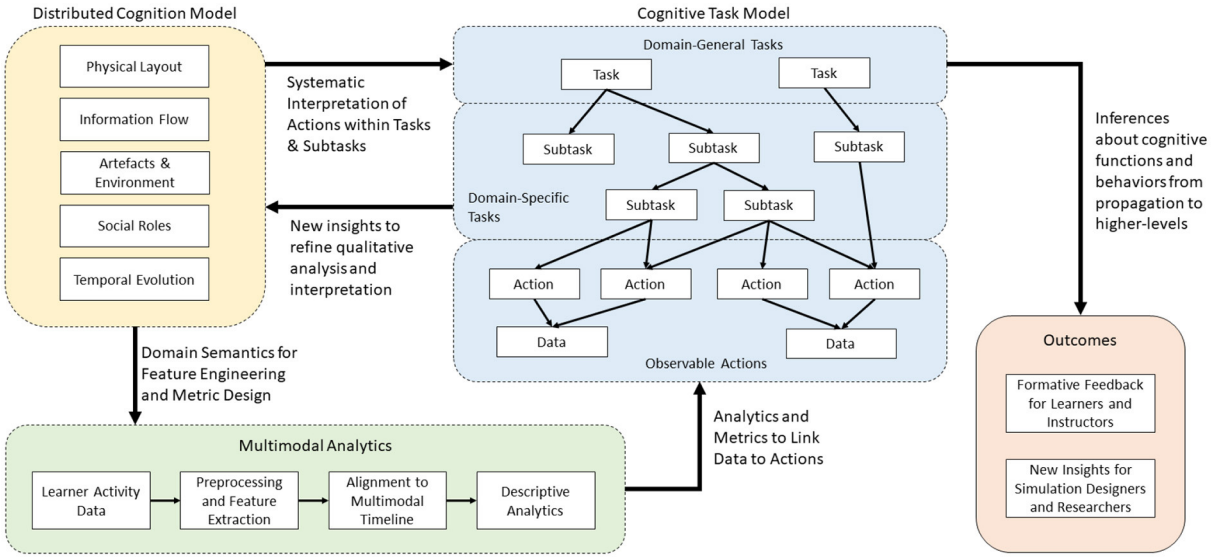

FIGURE 2
The overall theoretical framework to combine qualitative DiCoT analysis with quantitative multimodal analytics for understanding learner behaviors in simulation-based training.

to determine when the nurses move between each of these four semantic regions (see Section 4.4.1). As a second example, in our nursing domain, the DiCoT analysis revealed the various artifacts that are semantically important to information flow (see Section 4.3.2). We can adopt this result by using the eye-tracking gaze data and mapping the raw x-y gaze position data onto instances where the nurse is looking at each of the semantically important artifacts identified by the DiCoT analysis (see Section 4.4.4). In this way, we use the results of DiCoT analysis to create algorithmic models that convert raw data (e.g., video, audio, etc.) into action- and behavior-level interpretations.

Once we convert from the raw data to the action- and behavior-level interpretations, they are mapped onto a common *timeline*. As a next step, we can develop algorithms to interpret temporal sequences of actions and behaviors, and roll them up into upper sub-task levels. Some actions only contribute to a single sub-task, but others may link to multiple sub-tasks. These multiple hierarchical links in the task model add expressivity to our task models, but may make the analysis process more challenging because multiple inferences may have to be made on similar action sequences using additional context information.

We systematize this interpretation process by once again introducing results from the qualitative DiCoT analysis of the task environment, as illustrated by the arrow linking distributed cognition (yellow) to the cognitive task model (blue) in Figure 2. Results from the DiCoT analysis can provide semantic context to the interpretation of learner actions within the environment, and map them onto the sub-tasks to which the individual action may contribute. For example, when analyzing a group of participants in a restaurant, collected sensor data, such as video analysis or accelerometers, may indicate that a specific participant was performing the action of cutting with a knife. This action may contribute to at least two potential disjoint sub-tasks of interest: eating food or cooking food. However, based on a previous DiCoT analysis of the environment, we know that the physical layout of the restaurant strongly mediates the interpretation of these two sub-tasks; cooking activities occur in the kitchen, while eating activities occur primarily in the dining room. By adding this semantic context derived from the physical layout theme of the DiCoT analysis, we know that we can simply look at the participant's position in the restaurant to disambiguate this knife cutting action. As an extension, if we captured participant dialog and additional video around the cutting event, we may use information flow DiCoT theme to analyze the motivations for this action within a given sub-task, for example, to deduce that one participant was dividing his portion of food to share with another as part of the eating process.

While this restaurant scenario analysis represents a simplistic example, it demonstrates the second way in which DiCoT is important for adding semantic context to our computational analysis. First, DiCoT informs the design of algorithms and models to convert raw data to action-level interpretations. Second, DiCoT provides context-specific disambiguation when mapping lower-level action and sub-tasks onto high-level tasks and sub-tasks. By iterative analysis, we can propagate learners' activities up to the highest-levels of the task

model to understand their cognitive, psychomotor, affective, and collaborative behaviors.

By presenting the learners and instructors with quantitative metrics and qualitative descriptions of learner activities at multiple levels of the task model hierarchy, we can provide a basis for further discussion at different levels of detail during simulation debrief, while also tracking progress and changes in learner behavior over time. In addition, the results generated from this computational analysis also provide additional insights back into semantic models of the domain and inform a richer qualitative (DiCoT) analysis and task model construction. This idea is illustrated by the cyclic link from the cognitive task model (blue) to distributed cognition (yellow) in Figure 2. For example, analysis of the data might reveal certain learner behaviors that are not well accounted for in the current DiCoT analysis and task models. By presenting this result back to researchers, these analysis models can be refined and updated to contain a more complete understanding of the task environment and learner behaviors. This creates the loop back in our framework. Qualitative DiCoT and task analysis methods provide domain semantics and systematic methods for interpreting collected learner data, and the analysis of collected learner data reveals new insights that can be used to refine the DiCoT and task models. In the next section, we apply our task modeling framework combined with the DiCoT and multimodal analyzes to our MRMB nurse training case-study.

4. Methods

In this section, we demonstrate application of our theoretical framework to a small case study of nurses training in an MRMB environment. We begin with a complete description of the case study, including description of the affordances of the simulation environment and all of the data that was collected for the analysis. After this, we show how each of the three components of our theoretical framework apply to interpreting and analyzing nurses' activities and behaviors in this domain. First, we explain the construction and structure of the complete cognitive task model, from the high-level abstract cognitive tasks down to specific actions and observable data. Second, we describe a DiCoT analysis of the training environment, explaining each of the five themes in depth. Finally, we present a computational architecture, based on multimodal analysis, which tracks the raw multimodal data collected from the training environment through the cognitive task model to generate inferences, analytics, and performance metrics that describe the nurses' training behaviors within the context of the distributed cognition system.

Following the description of each component of the theoretical framework applied to the case study, we demonstrate the processes of following the collected data through the framework to generate inferences about nurse behaviors.

4.1. Case study-MRMB nurse training

The approach presented in this paper is supported by a case study that analyzes student nurses training in an MRMB environment. The training took place in a simulated hospital room, which was equipped with standard medical equipment and monitors for information display and communication of the providers orders. The patient was represented by a high-fidelity manikin that was exhibiting distress symptoms and a deteriorating health state. The simulated hospital room is displayed in Figure 3. All of the participating students were undergraduate (BSN) level nursing students in their first year and prior to the study had completed one semester of coursework, which included some simulations similar to those studied in this work. The simulations we study in this paper were part of the students' normal coursework requirements, and no changes to the content of the simulations were made by the researchers.

In more detail, the patient manikin is a SimMan 3G advanced patient simulator from Laerdal Healthcare that supports hands-on deliberate practice, development of decision-making skills, and improved communication and teamwork among learners (Laerdal Medical, 2022b). Prior to beginning the training, the basic scenario and simulation is pre-programmed using the Laerdal Learning Application (LLEAP) (Laerdal Medical, 2022a). This allows the instructors and simulation designers to set the initial state (vital signs, physical presentation, eye and chest movements, etc.) of the manikin, as well as a preset timeline of cue-action associations to change the state of the manikin as time progresses and the scenario evolves. For example, the timeline might be programmed to make the manikin's heart rate rise steadily if a nurse does not begin to administer proper medication within 10 min of the start of the training episode.

In addition to these presets created prior to training, an instructor in a control room can modify the patient state in real-time by interacting through the LLEAP software. The instructors watch the simulation from behind a one-way glass partition, allowing them to observe the nurses' activities, conversations, and interventions. Then, based on the nurses' specific actions (or lack of actions), the instructor makes real-time modifications to the simulation on the LLEAP software. The instructor can also talk as the patient through a microphone in the control room, which can be heard through speakers on the manikin. In the current study, which represented an early training exercise in the nursing curriculum, the instructor was closely involved in the progression of the simulation and manikin.

Three groups of eight nursing students participated in the study over 2 days, taking turns playing their assigned roles in each scenario. The primary participant in each instance of the simulation was a nurse performing a routine assessment on a hospital patient, and discovering a condition that required immediate attention and additional interventions.

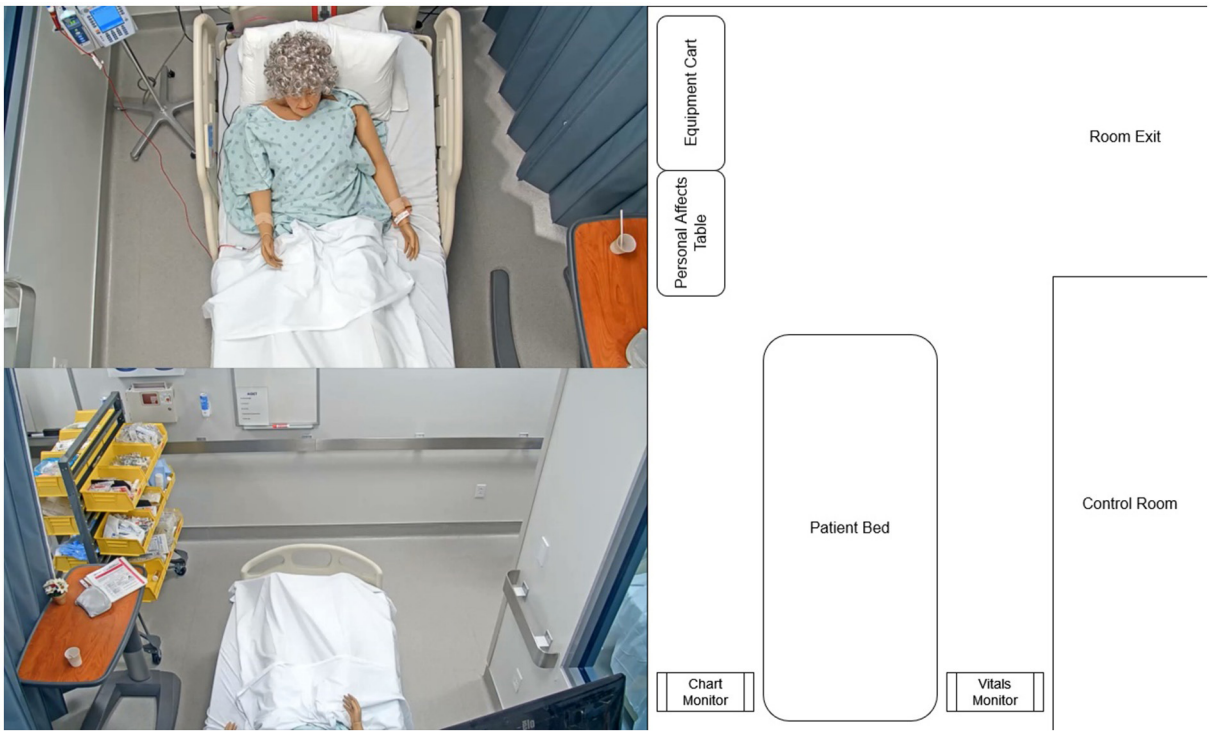

FIGURE 3
Layout of the simulated hospital room from three viewpoints: the head camera (top-left), foot camera (bottom-left), and an abstract map representation (right).

After diagnosing the patient's condition and performing any relevant immediate stabilization, the nurse was required to call the patient's assigned medical provider to confirm an intervention that would alleviate the patient's newly discovered condition.

Students in the group who were not actively participating in a given run of the scenario watched from a live camera feed in a separate debriefing room. After each scenario was completed, the instructors and the participants joined the full group in the debriefing room, and the instructor guided a discussion-based debriefing of the simulation. Each instance of the simulation took between 5 and 20 min, and parameters of the patient's condition were changed between each run to ensure the next set of students did not come into the scenario with full knowledge of the condition and the required intervention.

All students who participated in the study provided their informed consent. With this consent, we collected data using multiple sensors: (1) video data from two overhead cameras that captured the physical movement and activities of all agents in the room (nurses, providers, and the manikin); (2) audio data also from the camera videos that captured the nurse's dialog with the patient and the provider; (3) the simulation log files that tracked all of the patient's vital signs and data from the sensors on the manikin. In addition, a few students, who provided a second informed consent on collecting eye tracking data, wore eye tracking glasses that allowed us to record their gaze as they worked through the simulated scenario. The study was approved by the Vanderbilt University Institutional Review Board (IRB).

In this paper, we chose two of the recorded scenarios for our case study, in both of which the primary participant wore the eye tracking glasses. In the first scenario (S1), the fictitious patient, Patrice Davis, is receiving an infusion of blood after a bowel resection surgery the night prior. The patient called the nurse stating that she is not feeling well. The goal for this training scenario is for the nurse to assess the patient and diagnose that the wrong blood type is being administered to the patient. The intervention requires the nurse to stop the current infusion and call the provider to discuss further treatment. The primary participant in S1 was a 23 year old female nursing student.

In the second scenario (S2), the same fictitious patient, sometime later in the day, again calls the nurse complaining of pain in the right leg, stating that yesterday "it wasn't bothering me that much but today the pain is worse." The goal of this training exercise is for the nurse to assess the patient and diagnose a potential deep-vein thrombosis (blood clot) in her

right leg. The intervention requires the nurse to call the provider for updated treatment orders and to schedule medical imaging for the patient. The primary participant in S2 was a 24 year old female nursing student.

4.2. Cognitive task analysis for learner behaviors

Using the cognitive task analysis methods previously described, we generated a comprehensive task hierarchy for the nurse training domain. This hierarchy is illustrated in Figure 4. At the highest level of the task model, the overall task breaks down into three primary subtasks: (1) *Information gathering*, (2) *Assessment*, and (3) *Intervention*.

Information gathering represents the processes nurses apply to retrieve new information and monitor ongoing concepts. This process can be further characterized as either *general* or *diagnostic*. In general information gathering, nurses collect non-specific patient and situational information that they use to generate an overall mental model of the patient state. The information collected in this phase is largely standardized for each patient; for example, vital signs are often collected to give a broad overview of patient health. The mental model generated during this phase then leads the nurse to the diagnostic information gathering phase, where the nurse collects more pointed and specific information in service of diagnosing a specific issue with the patient. For example, if dialogue during the general information gathering phase reveals that the patient is experiencing leg pain, then the nurse might follow-up with a physical examination of the leg during the diagnostic phase in order to gather more specific information about the issue.

Assessment represents the processes used to synthesize gathered information in order to construct and evaluate specific solutions and interventions. In addition, we further decompose assessment into intervention *construction* and intervention *evaluation*. During construction, nurses synthesize and combine the information gathered from the environment to generate an intervention that represents a plan of action(s). By drawing on their prior knowledge of patient health and clinical procedures, and their current mental model of this specific patient established from the gathered information, nurses differentially construct a plan for how to help the patient.

During evaluation, similar processes are applied to synthesize information, but this time with a further emphasis placed on monitoring the progress of patient health over time. Temporally, the evaluation phase typically takes place after the nurse has already intervened in some way, and serves as a method to verify that progress toward the intervention goals is being achieved. The evaluation results in one of two possibilities depending on whether progress is made: either continue the intervention further or stop the intervention and re-assess to establish a new plan.

Intervention represents the actions and processes that nurses take in service of a specific goal related to patient health. These interventions are characterized as either *stabilization* or *treatment* procedures. During stabilization, the goal of the nurse is to fix any immediate threats to patient health. For example, in scenario S1 of our case study, the nurse typically turns off the infusion of blood, so that no further harm comes to the patient because of the incorrect blood type infusion. This action does not actually solve the underlying problem, i.e., the patient requires a different blood type infusion, but rather represents mitigation of an immediate threat before treatment of the underlying problem can begin. As a second example, if a patient were to stop breathing, the nurse would typically start resuscitation procedures. Here again, these resuscitation procedures do not fix the underlying cause of the patient's condition, but rather stabilizes the patient back to a point where they are not in immediate danger so that treatment of the underlying condition can begin.

In the treatment phase of intervention, the nurses' actions are in service of fixing underlying health issues that could cause danger to the patient's health in the future. For example, a nurse might start administration of chemotherapy drugs for a cancer patient. In this case, the medication is not designed to help immediate symptoms, but is rather part of a longer term plan to fix the underlying condition and put the cancer into remission. During the treatment phase, nurses will either start/continue an existing treatment order if they are aware of the patient's condition and a provider has prescribed the treatment. If the nurse finds a new condition in the patient, they will contact a provider to follow-up and get a new treatment order.

4.3. DiCoT analysis

As discussed, the DiCoT framework with its five themes: (1) physical layout, (2) artifacts and environment, (3) social structures, (4) information flow, and (5) temporal evolution; provides a qualitative framework for analyzing learner activities in the training environment. Results from this qualitative analysis then provides the basis for analyzing the multimodal data and inferring nurse activity and behavior information with supporting context. Figure 5 illustrates this idea in context. In this example, the nurse distributes her cognition across all five of the themes:

1. Physical layout by her position on the left and right sides of the bed;
2. artifacts and environment by her physical interactions with the available instrumentation and patient manikin;
3. Social roles by her verbal communication with the patient manikin;

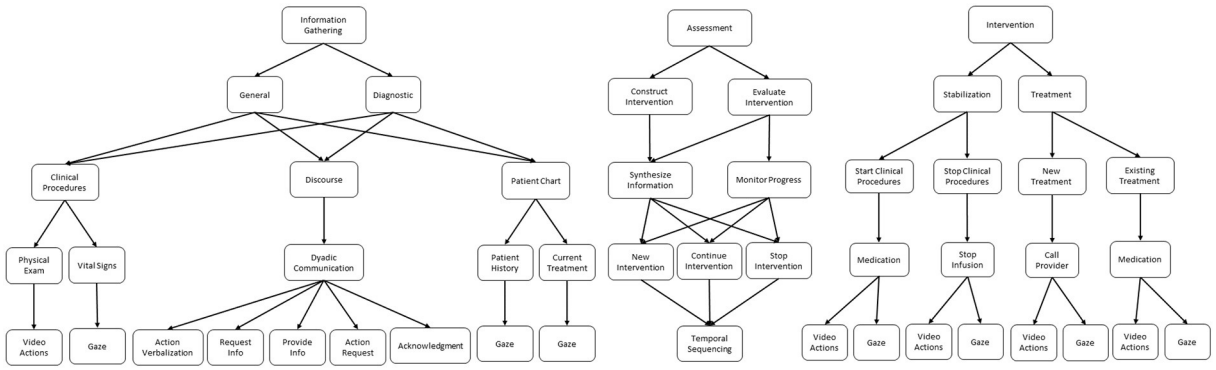

FIGURE 4
Cognitive task model for the nursing simulation domain.

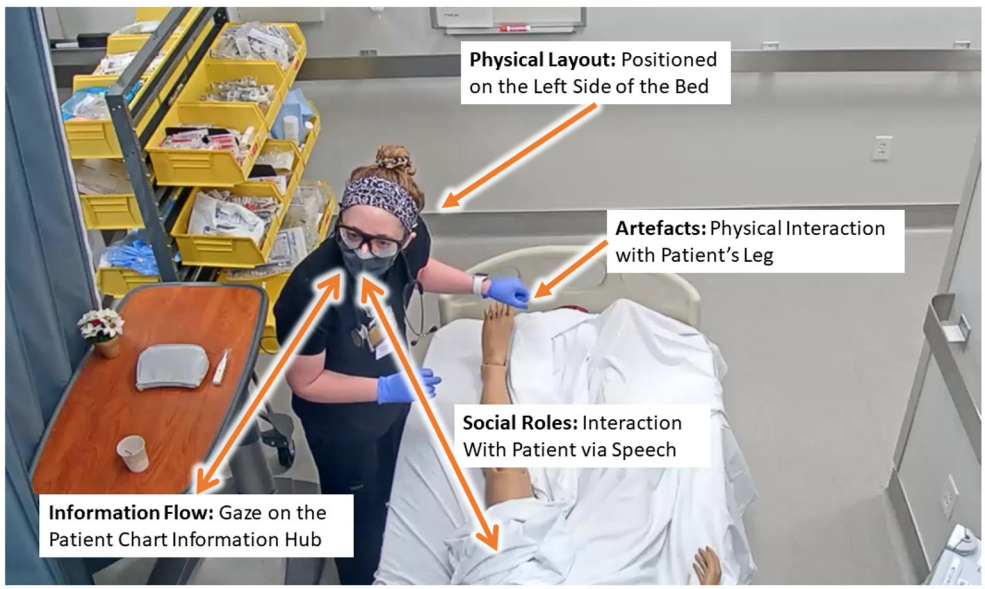

FIGURE 5
Example of the distributed cognition in the context of nurse training across the physical layout, artifacts, and social themes.

4. Information flow by her referencing of the patient chart monitor; and
5. Temporal evolution by following the sequence of her actions in the environment over time.

Using the five themes and 18 principles (see Table 1), we performed a DiCoT analysis of our nurse training simulation scenarios. We discuss our analysis for each the five themes in greater detail next. Similar to the analysis in Rybing et al. (2016), references to the specific principles are listed parenthetically as they relate to the description of each theme. For example, (P1) refers to Principle 1, i.e., Space and Cognition.

4.3.1. Physical layout theme

The complete layout of the room from three viewpoints can be seen in Figure 3. For the remainder of the paper, when discussing physical positions we will describe the positions in reference to the map view shown on the right-hand side of this figure. For example, *left of the bed* describes the area on the left-hand side of the map containing the patient chart and personal effects tray, while *foot of the bed* describes the area at the top of the map containing the equipment cart and the doorway.

The overall physical layout covered in the simulation environment mimics the layout of a typical hospital room, where the trained nurses apply their learned skills on real patients (P3,

P17). In the center of the room along the back wall is the patient bed, where the manikin is placed. To the right of the bed is a computer monitor that displays the vital signs of the patient as graphs (P2, P7). The default graphs and other vital displays are large enough so that the nurse can see them from any position in the room (P5, P6, P7), but the nurse can physically interact with the monitor to test certain vital signs and to get more information when she is on the right side of the bed (P1, P5, P6, P7). To the left of the bed is a second computer monitor that displays the patient's chart. The information on this chart is displayed in smaller text font, so the nurse has to be close to the screen to read information and needs to scroll on the screen to view all of the information. In other words, the nurse must move to the left side of the bed to access this chart (P1, P5, P6, P7). Past the foot of the bed, the room opens into a larger area that contains a cart of medical supplies that may be needed to perform clinical procedures (e.g., gloves, masks, needles, tubing, etc.) (P7). Finally, outside of the room is a medication dispensary; the nurses must leave the room and walk to the dispensary to retrieve needed patient medications (P5, P6, P7).

Given the physical arrangement of the room, we divided the physical space of the simulation into four regions that nurses may move between: (1) left of the bed, (2) right of the bed, (3) foot of the bed, and (4) outside the room. As discussed, each of these regions has available equipment and information that the nurses can use to accomplish their goals. Therefore, they may have to move between the regions to achieve specific goals. At the right side of the bed, nurses can perform clinical procedures, such as taking vital signs or interacting with other stationary equipment (e.g., IV pump, oxygen unit). These clinical procedures are components of the *information gathering* or *intervention* tasks in the cognitive task model.

At the left side of the bed, nurses can primarily perform *information gathering* tasks, such as looking at the patient chart or using the phone in the room to call medical providers. However, when on the left side of the bed, nurses may also cross-reference information from the vitals monitor that is on the right side of the bed (P1). This sort of cross-referencing is often accompanied by subtle body movements, for example, deictic gestures that involve pointing at the screen (P4).

The foot of the bed acts as a transition area for high-level cognitive tasks and lower-level sub-tasks. The training nurses enter the room through this area, establish their current goals, their observation (P6) and their situational awareness (P7) in relation the patient in the room. The nurses pass through this region when moving from the left side of the bed to the right (and vice-versa), while gathering information and making decisions on what clinical procedures to perform (P1). They often pick up equipment from the cart along the way (P7). Nurses also have to pass through the foot of the bed to visit the medication cart, or otherwise exit the room. When doing so, the foot of the bed provides a final moment of situation awareness before their horizon of observation shifts and they are no longer directly viewing the room (P6, P7).

4.3.2. Artifacts and environment theme

Within the simulation environment, the actors, in particular the nurses, utilize a variety of artifacts to support their training activities that are outlined in our task model. The first set of artifacts comes primarily in the form of medical equipment; some of them appear in Figure 3, and several have been discussed in previous sections of this paper. This equipment is designed to mimic the look and feel of a real hospital room, serving the primary goal of the simulation to gain transferable skills (P17), while also providing an interface into the patient data and a means for conducting procedures on the patient. Therefore, the medical equipment serve primarily as mediating artifacts (P15), which transform measurements, such as the vital signs of the patient into textual and graphical information that can be interpreted by the nurses (P9, P15).

Another important artifact in the simulation is the script, which is a set of guidelines set by the instructor about the unfolding of events in the scenario. The script outlines the initial conditions (e.g., the patient's condition, expected vitals at start), as well as a set of behavioral triggers (P14) for how the scenario should evolve given the potential actions (or lack of actions) performed by the nurse. For example, the script might specify that if the nurse does not begin infusing medication within 3 min after the scenario begins, the patient's blood pressure will drop. These scripts' trigger factors mediate the temporal evolution of the simulation (P15, see Section 4.3.5)

The manikin, representing the human patient, is another important artifact for the simulation. It provides an interface for the instructor to construct and guide the evolution of the scenario. The manikin is programmable; therefore, the instructor can digitally set parameters for the patient manikin (e.g., vital signs, movements, and conversations), which are then physically enacted by the manikin system (P13). During dialogue between the nurse and patient, the instructor speaks as the patient through the manikin offering additional information to the nurses (P10), as well as instructional scaffolding (P16), when needed. For example, if the nurse fails to take the patient's temperature, the instructor might scaffold this behavior by making a remark through the patient, such as "I also feel a chill," which may prompt the nurse to check for a fever by taking the patient's temperature. These dialogue acts can also be used by the instructor to evaluate the nurse's understanding and thought processes. For example, consider the dialogue sequence from S1 shown in Table 2. In this case, the nurse has concluded that the blood transfusion is causing the patient's issues, but in order to verify the nurse's understanding, the instructor asks a clarifying question through the manikin.

TABLE 2 Sample dialogue from S1 demonstrating evaluation of the nurse.

1	Nurse:	I'm going to stop this infusion really quickly.
2	Patient:	Why?
3	Nurse:	Because when we give red blood cells, an indication that you're having a reaction to it is low back pain and feeling itchy. So it sounds like you're having a reaction to the blood transfusion.

4.3.3. Social structures theme

Within SBT, there are three main types of users (Rybing, 2018). First, learners (or participants) represent those who participate in the simulation with the purpose of learning skills or having their performance evaluated (Meakim et al., 2013). Second, instructors (or teachers) are those who participate in the simulation with the purpose of directing the simulation to produce learning outcomes for the learners (Meakim et al., 2013). Finally, confederates (or embedded participants) are those who participate in the simulation with the purpose of enabling or guiding the scenario in some way (Meakim et al., 2013). The social structures of the simulation can be derived from the three basic user types of SBT.

In our nursing case-study, each instance of the simulation has three basic users: two students and the instructor. The students act as learners in the simulation, one taking the role of the primary nurse and one taking the role of the medical provider. The instructor takes a dual role as both the teacher as well as a confederate playing the part of the patient. The patient is enacted through the manikin mediating artifact described in the previous section.

4.3.4. Information flow theme

The primary goal for the nurse training in the MRMB simulation is to collect sufficient information about the patient (i.e., the *information-gathering* task) in order to make a diagnosis of the patient's condition (i.e., the *assessment* task). Then, the nurse has to act to alleviate the patient's discomfort and attempt to improve their health state; this is the *intervention* task. Thus, the movement (P8) and transformation (i.e., interpretation) (P9) of information is critical to making the correct diagnosis and conducting the right intervention. There are four primary sources of information in the simulation that follows the general structure of the *information-gathering* sub-tasks in the task model (Figure 4).

The first information source is the primary nurse, who typically provides information in the form of clinical knowledge that is previously learned during schooling and from prior experiences. This clinical knowledge includes

- Declarative knowledge, e.g., what is the nominal range for blood pressure?
- Procedural knowledge, e.g., how does one measure blood pressure accurately?
- Inferred associations using prior knowledge and observed information, e.g., given that the measured blood pressure is greater than normal, does it explain the conditions that the patient is experiencing?
- Diagnostic inferences, e.g., What may be the cause(s)?

It is important to note that the above is considered to be prior information, and not included as an element of *information-gathering* in the task model. Instead it is looked upon as a fixed input to the simulation system. The nurse may be required to recall this knowledge during the simulation, but this recall may not require any form of enactment and interaction in terms of a specific information gathering task within the training scenario.

Next, the patient's electronic medical record (EMR), also known as the patient's chart, is an information source containing a comprehensive history of the patient's prior symptoms, conditions, and treatments. The chart acts primarily as an information hub (P10), which allows the nurse to quickly reference the patient's history in a comprehensive way. However, it also plays the role of a mediating artifact (P15), since the chart is generally divided into sections allowing the nurse to access the relevant historical information related to the current diagnosis task. Additionally, since the chart contains notes from previous nurse shifts and the treatment being currently administered to the patient, it also helps the nurse trainee to better analyze the patient's trajectory and current condition, and use this to determine their goals and the tasks they need to perform (P17).

Third, the nurse is able to perform clinical procedures on the manikin and gather information about the patient's health conditions. These clinical procedures take a variety of forms, but the most common is collecting and characterizing the patient's vital signs. Nurses make use of the clinical equipment as mediating artifacts (P15) to make measurements on the patient and assess their condition. The mediating artifacts transform measurements into textual and graphical information for easier interpretation by nurses and other providers (P9). The output information is aggregated and displayed on the vitals monitor (see Section 4.3.1), which then acts as an information hub (P10). Other clinical procedures can also be performed by the nurses as needed. For example, if a patient is having pain in one of their legs, as in S2, the nurse might perform a physical examination of the patient's leg to gain more information.

Finally, social interactions between the nurse and the patient provide important information that is not measured directly. The instructor speaks through the patient to provide some of this information to the nurse(s). This information often provides elaborations of the patient's symptoms and additional symptoms that are not directly measured. For example, the patient might describe the location, severity, and history of their

pain. These social interaction represent the *discourse* sub-task in the task model.

As the simulated scenario evolves, information primarily flows from the four information sources described above to the nurse (P8), who then process the information (P9, P18) and act on it (P14). When nurses enters the room, they generally begin with a brief interaction with the patient, and this results in information transfer about the patient's general conditions and symptoms from the patient to the nurses. This typically provides an initial baseline for the nurses to check for additional symptoms and start making diagnostic inferences (P13). Thus, it is a component of the *general information gathering* sub-task in the task model.

Next, the nurses typically take some time to reference and review the chart, synthesizing the information that they just heard with the patient history before returning to a more extended dialog with the patient to extract more specific information to support diagnostic inferences. The nurses may ask a series of questions to the patient combining what they saw in the chart with their clinical knowledge (P14). This discussion is typically followed-up by one or more clinical procedures, such as taking vital signs and performing physical examinations. This cycle of discussion with the patient followed by clinical procedures can then be repeated as necessary until the nurse reaches some form of conclusion about the patient's condition. At a higher-level, this can also be thought of as a cycle between the *diagnostic information gathering* and the *synthesize information and construct intervention* sub-tasks in the task model.

Up to this point in the simulation, nearly all of the information has been flowing in from the other information sources in the environment to the nurses (P8). However, once nurses collect sufficient information to reach a conclusion, the process reverses and the synthesized information and resulting conclusions are provided back to the rest of the system through their resulting actions. Common actions at this point include explaining the situation to the patient, starting and stopping certain treatments (e.g., medications), and calling the medical provider to give an update and request updated treatment. These actions and the general flow of information from the nurse to the environment is an enactment of the *intervention* task in the task model.

4.3.5. Temporal evolution theme

The simulation evolves over time in one of two possible ways: through nurse actions or nurse inaction. The instructor has a script artifact which outlines a set of behavioral triggers that detail how the scenario should change (P14). Most of this script deals primarily with triggers due to nurse inaction. For example, the script might dictate that if the nurse does not start medication within 5 min of the scenario starting, then the patient's heart rate begins to climb steadily. On the other hand, scenario changes due to nurse actions are primarily dictated by medical and social responses based on the judgement of the instructor (P3). The nurses gather information to evaluate the situation. Then, based on their evaluations, the nurses intervene to alleviate the patient's conditions. Based on that intervention (or lack thereof), the instructor modifies the scenario. If the intervention was correct, then the patient improves and the simulation ends, but if the intervention was incorrect, then the instructor may further decline the patient's health and the nurse must re-evaluate the presented information and try a new intervention strategy. The temporal evolution of the simulation is built primarily along this cycle of information gathering and intervention.

4.4. Computational framework

One of the primary goals of this work is to show how quantitative data can enhance the qualitative DiCoT analysis and integrate this analysis with task modeling framework to better analyze and interpret learner behaviors. To do this, we create a computational framework that takes the raw data collected from the different sensors, maps it onto specific features derived from the DiCoT analysis and then interprets them using the task hierarchy. In our case study, we perform analysis on two raw data sources,

1. Overhead video cameras, and
2. Eye tracking glasses.

These map onto four feature modalities that form the basis of our analyzes: (1) position, (2) action, (3) speech, and (4) gaze. The complete computational framework is illustrated as a block-diagram in Figure 6.

From a combination of the four feature modalities, we construct a complete progression of activities and events on a *timeline*. A complete timeline for the case study analysis of scenario 1 is shown in Figure 9, and a similar timeline for scenario 2 is illustrated in Figure 10. This timeline structure forms the basis for a second level of analyzes, where information from the extracted features across the different modalities are combined to extract patterns. By combining the aligned features, extracted across the different modalities, we can extract activity information in context, and propagate the low-level actions up the task model to generate inferences about the nurse(s) cognitive processes and their behaviors. We provide a descriptive account for our analyzes of each of the modalities in the subsections that follow.

4.4.1. Position modality

The nurse's positions in the simulated hospital room are derived using visual object motion tracking techniques applied to the video from the two overhead cameras. Our motion

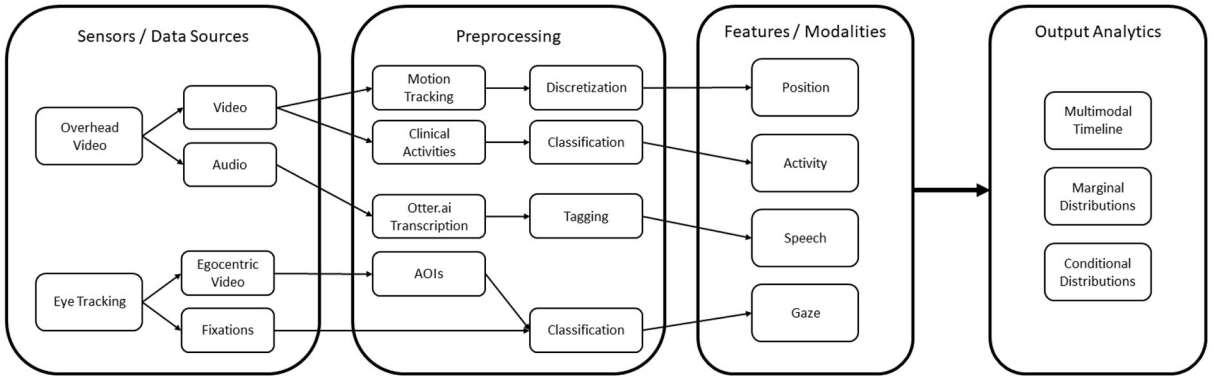

FIGURE 6
The overall computational architecture used for the quantitative analysis.

tracking techniques are derived from the tracking-by-detection paradigm, which is a two stage approach to tracking Sun et al. (2020). First, in each frame of video, deep learning-based object detection models localize people that appear in the video frame and represent them with bounding boxes. After this detection step, the detections are merged together frame-by-frame into a timeline based on a matching algorithm.

In our case studies, we use the matching cascade algorithm originally developed in Wojke et al. (2017), and later refined for static cameras by Fu et al. (2019). The matching cascade algorithm matches bounding boxes and tracks between subsequent frames based on the distance between the two bounding boxes and approximation of the velocity of the object in the track. In addition, the matching cascade algorithm matches the bounding boxes iteratively based on the age of a detection and track, leading to lower false positive rates.

However, these motion tracking techniques only produce a track of the nurses in reference to the video frame. We need to map these tracks into the nurses' positions in the hospital room as we have described in the physical layout theme of our DiCoT framework. To accomplish this, we extend our traditional motion tracking techniques to project the camera-space motion tracks onto a top-down map representation of the environment (see Figure 3, Right).

Our approach for mapping these camera-space tracks onto this hospital room space computes a planar homography, which associates known points in the camera-space to known points in the map-space using rotation, translation, and scaling operators. Given the computed homography matrix, we can project the camera-space tracks onto the room-space for each frame of video, using the center of the person's bounding box as the projected point. This results in a continuous time-series of nurse positions in the simulation room relative to the top-down map. Further details of this map-projection object tracking can be found in Vatral et al. (2021).

While the continuous time-series of nurse positions in the hospital room is a useful analysis tool, on it own, it lacks the semantic context necessary for meaningful insights. To add this semantic context back to the position data, we discretize the continuous positions into four regions developed using DiCoT analysis (see Section 4.3.1): (1) left of the bed, (2) right of bed, (3) foot of the bed, and (4) outside the room. To perform this discretization, we define a polygonal region on the top-down map of the hospital room for each of the DiCoT semantic regions. Then for each timestamp of the continuous track, we check the polygonal region that contains the nurse's position and assign that label to the given timestamp. This allows us to track in terms of time intervals of nurse positions in the different semantic regions of the room, and when they transition between these regions.

4.4.2. Action modality

In addition to providing position information, analysis of the overhead camera video also provides important information and context for the actions that the nurse performs in the training scenario. Specifically for this case study, we annotate instances in the video where the nurse performs an action by physically interacting with any of the artifacts in the MRMB environment previously identified from the DiCoT analysis.

Additional contextual information can be derived by combining the physical activity that defines an action with other modalities. For example, analyzing speech (see Section 4.4.3) may provide additional information about why a nurse is performing a specific action, or how two nurses are coordinating their actions, for example, when they are jointly performing a procedure. Similarly, a coding of the nurses' gaze (see Section 4.4.4) may provide additional information about how a nurse is performing an action. In some situations, the nurse may look at the same object that they are physically interacting with; in

other situations, the nurse may look at a different object than the one they are physically interacting with. As an example, while physically examining a patient, a nurse may turn their gaze to the vitals monitor to see how their current measurement may match with other vital signs (e.g., blood pressure being measured and heart rate of the patient). These examples clearly illustrate the importance of combining information across modalities for action annotation to gain a complete understanding of the nurses' activities in the training environment.

To perform action annotation, we have developed a coding schema based on the artifacts from the DiCoT analysis, which represents all of the high-level objects that nurses physically interact with during the simulation. These objects are primarily medical equipment, e.g., the patient chart, the vitals monitor, and the IV unit. They also include specific parts of the patient that are relevant for physical examination in these scenarios, e.g., the patient's hands, legs, body, and head. In total, we coded nurse actions into 13 categories for the two scenarios in our case studies, which can be seen on the timelines for each scenario (Figures 9, 10). The annotation recorded the action category along with start and end timestamps with a one-half second fidelity. Nurses were considered to be performing a given action category if they were physically interacting with the action object using some part of their body, typically their hands. For example, if the nurse was holding a phone or touching the dial pad, then they were coded as performing the *phone* action. Alternatively, if the nurse's hands were on the mouse and keyboard of the chart computer, then they were coded as performing the *patient-chart* action.

4.4.3. Speech modality

Raw speech is collected from multiple streams that include the audio from the two overhead cameras, and each of the Tobii eye tracking glasses. For this case study, we only analyzed audio from the overhead camera at the head of the bed. In future work, particularly during simulations with a greater focus on teamwork, we intend to analyze audio by creating an egocentric framework for each agent in the training scenario.

While raw recorded speech patterns are useful for some tasks (e.g., emotion detection), most NLP tasks perform analysis directly on a body of text, which requires raw audio to first be transcribed as a preprocessing task. For the current case study, we utilized the Otter.ai speech transcription service (Otter.ai, 2022). After transcription, the speech text is annotated (tagged) with specific events for analysis via the BRAT Rapid Annotation Tool (BRAT) (Stenetorp et al., 2012). Based on the task model (see Section 4.2), we developed a tagging schema for the speech data, which breaks down the dialogue into six speech event tags, which are enumerated below:

1. *Generic, introduction*: Refers to introductory speech such as greetings.
2. *Generic, acknowledgment*: Refers to generic acknowledgments of understanding, typically used as part of closed-loop communication patterns.
3. *Information, request*: Refers to the soliciting of information from another person.
4. *Information, provide*: Refers to the furnishing of information to another person.
5. *Action, verbalization*: Refers to the verbalization and explanation of an action. This verbalization can occur before an action begins, while an action is being performed, or after an action has been completed.
6. *Action, request*: Refers to a request for another person to perform an action, typically taking the form of either a question (e.g., Will you do this?) or a command (e.g., Do this).

Figure 7 illustrates a tagged speech snippet from Scenario 1. In this part of Scenario 1, the patient indicates that her "lower back hurts a little bit" and she feels "just kind of itchy all over." These are examples of the patient providing information to the nurse, so they are tagged as "Information, provide." The nurse then responds with an "Action, request" by indicating that she (the nurse) would like to check the patient's vitals. The nurse then asks the instructor whether the vital signs device is connected to the blood pressure cuff, which is tagged as another request for information. The instructor then responds to the nurse in the affirmative, which is another instance of "Information, provide." Notably, the nurse asks "was *that* connected to the blood pressure cuff?" The italicized "that" in this example is ambiguous if speech is the only modality considered for analysis. However, applying the vision and gaze modalities make it clear that the nurse is referring to the device used to actually take the patient's blood pressure. This is an example of how multimodal approaches can augment the information obtained from simulation-based learning environments.

Additionally, it is important to note that there are transcription errors in Figure 7. An important research consideration is whether or not to correct these errors before conducting the analysis. Human-in-the-loop transcription corrections provide the cleanest text to feed into the language model during analysis; however, there is a trade-off. Human-engineered text is expensive to generate time-wise, as manually correcting transcriptions involves reading every piece of a transcribed block of text. With large corpora, this is infeasible. Additionally, human-in-the-loop transcription correction precludes online analysis, as a human would first have to manually edit the transcription before it is used in a downstream task. Lastly, there can be an advantage to having a certain degree of noise in the data, as this can prevent language models from overfitting. Contemporary language models are traditionally trained on large corpora of canonical text. Because speech is rarely canonical, fine-tuning a language model to recognize spoken text is a challenge. However, this can often

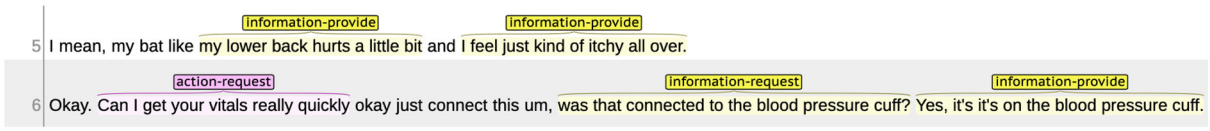

FIGURE 7
An example of dialogue from scenario 1 which has been annotated using the developed tagging schema.

be mitigated (at least in part) by injecting noise (misspellings, for example) in the data (Cochran et al., 2022). It is for these reasons we decided to annotate the text as-is from Otter.ai, without manually correcting the transcriptions.

4.4.4. Gaze modality

Gaze data is collected using Tobii Glasses 3. The glasses record multiple raw data streams including egocentric-view video, audio, eye gaze (2D and 3D), and inertial movement units (IMU) (Tobii Pro, 2022). The eye gaze data stream is sampled at 50 Hz and contains 2D coordinates corresponding to the egocentric video and 3D coordinates with respect to the camera's coordinate system. The egocentric video is sampled at 25 Hz in 1920x1080 resolution. The IMU sensors onboard the glasses include an accelerometer, gyroscope, and magnetometer, which are sampled at 100 Hz, 100 Hz, and 10 Hz, respectively. Through the combination of all data streams recorded by the Tobii glasses, the nurse's experience in the simulation is logged with high fidelity.

Given the high sampling rate and noise present in eye-tracking data, *fixation classification* is a common practice in the eye-tracking literature to pre-process raw gaze data and prepare it for further analysis (Bylinskii et al., 2015; Liu et al., 2018). Our initial pre-processing step applies Tobii's Velocity-Threshold Identification (I-VT) fixation filter to extract reliable fixation and saccade data. The classification algorithm identifies fixations and saccades based on the velocity of the eye's directional shift and a set of hyperparameters (Olsen, 2012). The default values provided by Tobii for the I-VT fixation filter were used during our analysis (Tobii Pro, 2012).

The final preprocessing step is to encode the fixation data into areas of interest (AOI) sequences. Linking fixations to AOIs bridges the gap between direct sensory output to domain-specific content, thus providing further insight into the nurses' attention and engagement. The temporal evolution of nurses' visual attention is represented by AOI sequences. In this study, AOI encoding from the fixation data is manually annotated to 11 objects of interest (OOI) that were selected based on the DiCoT analysis (patient, provider, screen chart, paper chart, vitals, medical tray, equipment, keyboard, instructor, one-way mirror, ground). Each of these physical objects are treated as an AOI and are annotated using the egocentric video. The manual tagging is performed through visual inspection of the egocentric video with fixation data overlaid. In each case where the nurse fixates on one of the AOIs, the start and end time of the fixations are recorded. An example of the gaze overlaid on the video is shown in Figure 8, where the red circle marks the fixation.

FIGURE 8
An example of fixation overlay from Scenario 2 used for manual annotation. In this frame, the resulting AOI is "patient".

5. Case-study analysis

The alignment and processing of multiple data modalities reveals new inferences about the simulation and the nurse's behaviors. In this section, we analyze and interpret these integrated multimodal timelines (Figures 9, 10) in depth for each scenario. We provide details of the basic breakdown of nurse actions and use the DiCoT framework to interpret these actions in context and map them onto the task analysis hierarchy. In addition, we compare across the two scenarios to see how the nurses differed in their cognition and use of environmental affordances in the MRMB environment.

5.1. Scenario 1

For scenario S1, the timeline breaks down into approximately five high-level segments. The first segment follows the general *information gathering* task established by the task analysis model in Section 4.2. During this segment, the nurse first enters the room and greets the patient, and then

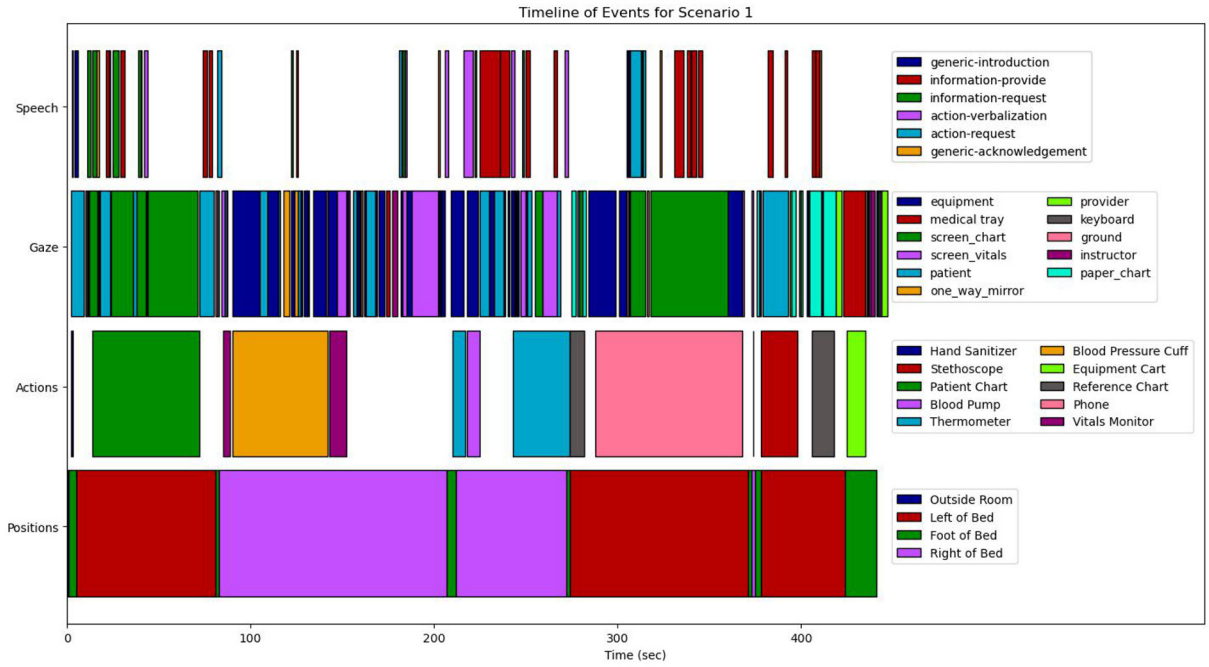

FIGURE 9
The complete timeline of events for scenario S1 containing annotated data from participant position, action, gaze, and speech.

moves off to the left side of the bed. In this position she begins a period of alternating between reading the patient chart and conversing with the patient, as indicated by her eye gaze moving between the patient and chart monitor. The conversation here is primarily pairs of *information-request* and *information-provide*, indicating that the nurse is asking the patient questions to clarify and expand on the information the nurse is reading from the chart.

Once the nurse decides she has enough information to build her initial mental model of this patient's situation, the simulation enters the second phase. This transition is marked by the nurse moving from the left side to the right side of the bed, as seen in the position modality around 80 s into the scenario. As previously shown from the DiCoT analysis, this movement between regions in the room is an important indicator of task transitions. During this new segment, the nurse moves to the *diagnostic information gathering* phase described in the task model. In this phase, the nurse increases her interactions with the equipment and the vitals monitor. We derive this information from the gaze, which shows movements between equipment, the vitals monitor, and the patient. In addition, her physical actions show interaction with the vitals monitor and the blood pressure cuff. In this segment, we can apply information from the DiCoT framework to provide additional context for establishing these action as diagnostic information gathering. Because of this movement from the left to the right side of the bed (physical layout) and the increased interaction with clinical equipment (artifacts and environment), we infer that the nurse is attempting to establish and refine her diagnostic inferences from the initial information gathering phase. She performs clinical procedures, such as taking additional vital sign measurements to aid her diagnostic hypothesis formation.

In this segment, we also see a reduction in dialogue, which likely has two causes. First, specific to this scenario, much of the information that can be obtained from the patient has already been gathered in the previous segment. Second, the cognitive load associated with performing clinical procedures (e.g., when taking a blood pressure reading) is likely higher than simply reading the patient chart. Because of this, the nurse may focus more on the clinical task at the expense of continuing conversations with the patient. This is especially true for novice trainee nurses who are still learning how to perform clinical procedures in correct and effective ways. Knowing that these clinical tasks require higher cognitive loads and having observed from the control room that the nurse reduced her dialogue, the instructor likely also intentionally reduced their conversations with the nurse during this period. The instructor may have spoken less through the patient while these tasks were being performed to avoid splitting the nurse's attention, conforming to the best practices during SBT (Fraser et al., 2015).

Around 220 s into the scenario, our video analysis shows that the nurse begins to interact with the blood pump, which implies a transition to the third segment of her overall task. According to the DiCoT analysis, the blood pump is a mediating artifact,

not an information source. Given this additional context, we can conclude that the nurse has reached the end of her diagnostic information gathering phase and has begun the *intervention* process in this new segment. Since the nurse interacts with the blood pump at the start of the intervention process, we can hypothesize that the nurse has reached a diagnostic conclusion, and suspects the blood infusion process. In other words, the nurse suspects that the patient is being administered the wrong blood type during infusion.

Specifically, in this segment the intervention represents the stabilization process, which requires the nurse to stop the blood infusion and prevent any further damage to the patient's health because of the infusion of the incorrect blood type. At the start of this segment, as our video analysis shows, the nurse stops the infusion, but the speech modality also records an *action-verbalization* event. The speech analysis module interprets the nurse telling the patient that she is turning off the blood infusion. This is immediately followed by the patient asking "Why?," and the nurse follows up with a proper diagnostic explanation, i.e., "an incorrect blood type is being infused." This discourse interaction, transcribed in Table 2, is an example of the dual social role of the instructor as both the *teacher* evaluating the nurse, and a *confederate* playing the part of the patient (see Section 4.3.3).

It is quite reasonable for a patient to ask questions about their condition and the treatments being administered in a real hospital setting. The instructor plays this role as the confederate. Indirectly, some of the questioning by the patient (i.e., the instructor as the confederate) also serves as an evaluation of the nurse who must explain her reasoning. This sort of evaluation questions arise from the instructor's role as the teacher, rather than the confederate. Since the instructor is playing both social roles, this discourse interaction may fulfill multiple pedagogical roles in the simulation scenario, i.e., how the nurse conveys diagnostic information to the patient to reassure them, and how the nurse has combined all of her observations to make diagnostic inferences. In this same time interval where the nurse interacts with the blood pump and verbally explains what she is doing, we also see her gaze moves between the equipment (the blood pump) and the patient, which is likely part of the social dynamics when interacting with a patient. The nurse should not ignore the patient while performing clinical procedures, which is exemplified here as the nurse shifting her gaze between the patient and the blood pump.

Once the stopping of the infusion is observed in our video analysis, the fourth segment of the simulation begins, with the transition marked again by the nurse's movement; this time the movement is from the right side of the bed back to the left side. This segment maps on to the *treatment intervention* phase of the task model. Since the diagnosed issue is not one that already has a physician prescribed treatment, the nurse calls the provider to make them aware of the new situation, as indicated by the *phone* action around 290 s into the scenario. We can see that during the period where the nurse is using the phone, her gaze is primarily on the chart monitor, likely because she is reading off the patient's information to the provider over the phone. This is also consistent with the speech acts, where we see several sequential *information-provide* acts, again likely because she is giving the patient's information to the provider over the phone.

During this same period, the speech also shows a few *action-request* events, which correspond with the nurse requesting that the provider come to the room to confirm the diagnosis. Shortly after the phone call, the final segment of the simulation begins, marked by the arrival of the provider around 390 s into the scenario. In this segment, we again see several sequential *information-provide* acts in the speech corresponding to the nurse explaining the patient's situation to the provider. The nurse and the provider then look at a reference chart, which contains information about the protocol to re-test blood type. Finally, the two move to the foot of the bed and begin examining the equipment cart, likely to collect the necessary equipment to draw the patient's blood. This marks the end of the training scenario, and the nurse moves on to a debrief session outside of the simulation hospital room.

5.2. Scenario 2

In scenario 2, the timeline breaks down into four high-level segments. Once again, the first segment represents the *general information gathering* task. The nurse enters the room and moves to the left of the bed by the patient chart monitor. During this initial movement period, there is a short sequence of alternating *information-request* and *information-provide* speech acts, indicating the nurse asking the patient initial questions to learn about their general background and current condition. Just as in S1, the initial movement to the left side of the bed is a significant indicator of entering an information gathering phase, as indicated by the physical layout DiCoT analysis.

This initial movement and speech is then followed by a long period of attention strictly on the chart monitor, as seen in both the actions and gaze, as well as the absence of any dialogue. As shown in the information flow DiCoT analysis, this chart monitor is a significant information hub in the room and further supports this segment as information gathering. The absence of dialogue here is also particularly interesting when compared to the nurse in scenario 1, who tended to multi-task dialogue with the patient while reading the chart. However, here we see a different information gathering strategy of first spending devoted time to the chart, followed by a shorter period of *information-request* and *information-provide* acts (e.g., question and answer) around 100 s into the scenario.

During this question and answer period, the nurse's position moves quickly between the foot of the bed, the right of the bed, and back to the left of the bed, with her gaze also moving rapidly between pieces of equipment and other artifacts in the room.

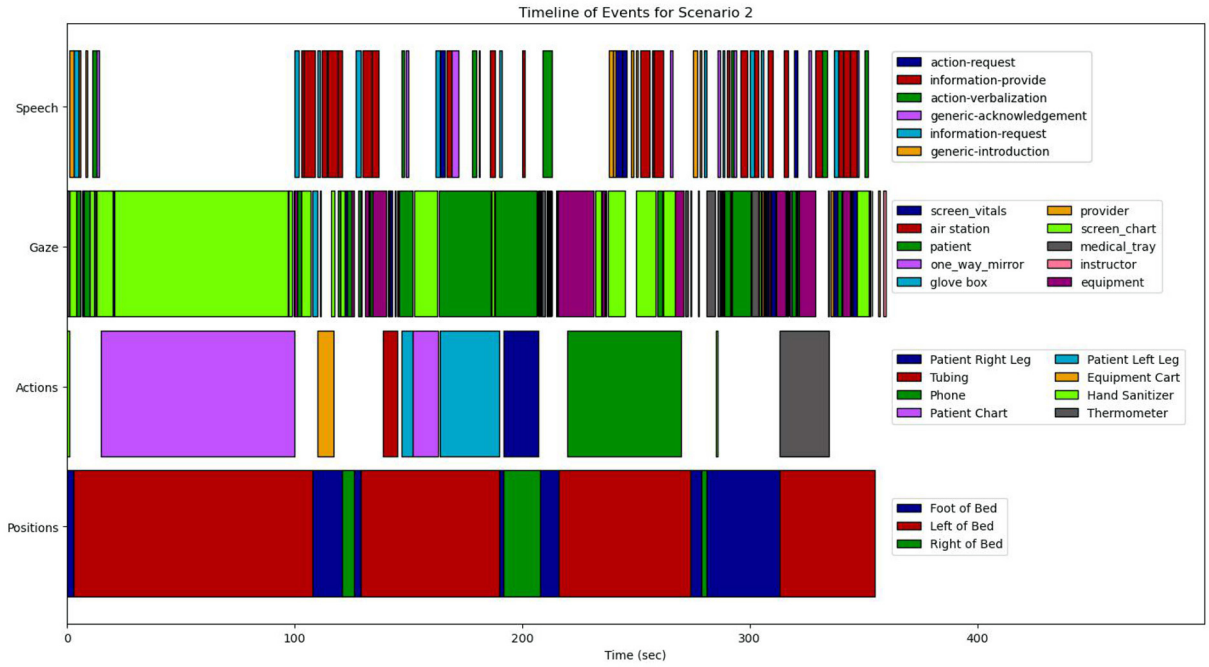

FIGURE 10
The complete timeline of events for scenario S2 containing annotated data from participant position, action, gaze, and speech.

On its own, it is unclear what exactly the purpose of these rapid movement and gaze changes are; however, given that this occurs while the dialogue is primarily question and answer, which is an information gathering task, it is likely that the movement and gaze are also related to the information gathering. While the nurse is using dialogue to gather information about the patient during this period, she is simultaneously also gathering information about the available equipment and physical layout of the room using her movement and gaze.

At this point, the second segment of the simulation begins, marked by the nurse moving back to the left side of the bed and her gaze now stabilizing back on the patient and chart, around 140 s into the scenario. Like scenario 1, the second segment represents *diagnostic information gathering*. Having determined patient history and the current issue with the patient, i.e., severe right leg pain, the nurse begins a *physical examination* of the patient in order to further refine her diagnosis of the problem.

The nurse begins examining the patient's left leg for a short period of time, while asking the patient whether certain areas that the nurse touches are tender. This is derived from our analysis of nurse's actions, which show physical interaction with the patient's leg, along with speech analysis which shows sequential *information-request* and *information-provide* acts. After this exchange, the nurse turns her gaze from the patient back to the chart, likely because she is surprised when the leg does not hurt to the touch. At this point, the information she obtained from dialogue with the patient and the patient chart does not match with the physical exam of the leg. Because of the conflicting information, the nurse looks back on the chart to recheck the information she previously gathered and her diagnostic hypothesis.

After a few more moments of examination and dialogue with the patient, the patient finally speaks up and says, "It's my other leg that hurts." At this point, the nurse quickly moves over to examine the right leg, as shown in the action data. There are several interesting points about this interaction. First, dialogue of the patient is another manifestation of the dual social role of the instructor. The instructor is acting as the patient in this moment, but also providing some instructional scaffolding, e.g., that the nurse needs to examine the other leg. By inhabiting this dual social role, the instructor can seamlessly introduce the instructional scaffolding into the simulation scenario by speaking through the patient.

Second, by combining data modalities, we gain a much deeper understanding of the nurse's activities in the training scenario. Because we have the eye gaze information and see that the nurse looks back at the chart, we interpret that the nurse realizes that there is an issue before being corrected by the patient. Pedagogically, this is important because it shows a level of metacognitive awareness in the nurse which we may not have realized otherwise. The nurse looks back on the chart to recheck her diagnostic hypothesis because of the conflicting information she has received that the patient's leg does not hurt to the touch. Without this gaze information, we may have surmised that the

nurse had gone down a wrong path, and would need to be corrected on her diagnostic hypothesis. However, her looking back to study the chart and asking questions to the patient made us realize through the analyzes that she was reconsidering her current diagnostic hypothesis.

After examining the right leg, the training scenario transitioned into the third segment, marked by the movement of the nurse from the right side of the bed where she was examining the leg back to the left side of the bed. This movement, around 220 s into the scenario, again highlights the physical layout theme of the DiCoT analysis. In this segment, the nurse began the *intervention process*. No stabilization processes are clinically necessary in this scenario, so the nurse immediately proceeded to *treatment*. Just as in S1, there was no physician prescribed treatment, so the nurse called the provider to update them and get a new treatment order, as indicated by the *phone* action. While on the phone, the nurse's gaze was primarily on the patient chart, with a few instances of looking back at the patient, and simultaneously her dialogue was a series of *information-provide* acts. This gaze and speech in combination indicate that she was reading patient information off the chart to the provider, and filling in additional details based on her observations and gathered information of the patient condition.

Shortly after the phone call, the scenario transitioned into the fourth segment, marked by the entry of the provider into the room and the nurse moving to the foot of the bed, around 290 s into the scenario. In this segment, the dialogue shows a series of sequential *information-request* and *information-provide* pairs, indicating that the provider was asking questions to the nurse and the nurse was answering based on her gathered information and assessment of the patient. During this sequence, the provider asked whether the nurse has gathered patient vitals. After realizing that she did not finish this task earlier, the nurse began to interact with the equipment to complete collecting the vital signs, as shown by the *thermometer* action and the nurse's gaze moving between equipment and the vitals screen. The scenario finished with a short discussion about the next steps for treatment, specifically the scheduling of a scan of the patient's leg, shown by the series of *information-provide* acts in the speech at the end of the timeline.

5.3. Cross-scenario discussion

In this section, we combine the analysis across both scenarios to demonstrate how the collected data supports the DiCoT analysis presented previously. For this analysis, we will focus on the three primary DiCoT themes which are typically analyzed: physical layout, information flow, and artifacts and environment. We will examine each of the three DiCoT themes individually and how the data-driven evidence supports the major conclusions from that theme.

To support the comparison between the contextually different scenarios, we computed a series of marginal and conditional distributions of the four data modalities. Figure 13 shows the marginal distribution of gaze across the entire scenario; Figure 11 shows the distribution of gaze conditioned on position in the room; and Figure 12 shows the distribution of speech conditioned on position in the room. These distributions were computed based on the modality-aligned timelines (Figures 9, 10) by dividing the sum of the time spent on a given modality class by the total scenario time. For example, to compute the percentage of equipment gaze events conditioned on being positioned on the left side of the bed, we divided the sum of the times spent looking at equipment while on the left side of the bed by the total time spent on the left side of the bed. By comparing the marginal and conditional distributions of the scenarios instead of the scenario timelines directly, we can help reduce the temporal autocorrelation caused by the differences between the scenario contexts. In other words, the distributions provide a more direct comparison between the two scenarios that does not care about the order in which nurses completed actions, since the order is highly dependent on the specific scenario and patient condition.

Beginning with the physical layout theme, a wealth of data supports the roles that space and physical layout play in the nurses' cognition. The timeline analysis shows that both nurses exhibit similar patterns in their movement through the physical space. Each nurse begins by entering the room through the door at the foot of the bed and immediately moving to the left side. The nurses stay on the left side to gather initial information from the chart and conversation with the patient before moving to the right side of the bed to begin their diagnostic clinical procedures. While the specifics of information gathering and clinical procedures differ between the two scenarios, the general movement patterns and associated tasks in these areas of the room remain very similar.

Support for the roles of these spaces can also be seen through the conditional distributions of gaze in Figure 11. For both nurses, the percentage of gaze events focused on the chart and the patient was higher when they were on the left side of the bed, while the percentage focused on the vitals screen was higher when they were on the right side of the bed. This was particularly evident for scenario 1, where focus on the chart and vitals when on the left side of the bed were 25 and 2.8%, respectively. It changed to 1.9 and 18.5%, respectively when they were on the right side of the bed.

For scenario 2, while the difference in gaze for the chart monitor was fairly small, changing from 22.8% on the left down to 20.0% on the right, the difference in gaze for the vitals monitor was still quite large, with 10.5% when on the left and jumping to 40.0% when on the right. These differences between the left and right sides of the bed was also supported by the speech analysis. As shown in Figure 12, the nurses in both scenarios performed most of their dialogue

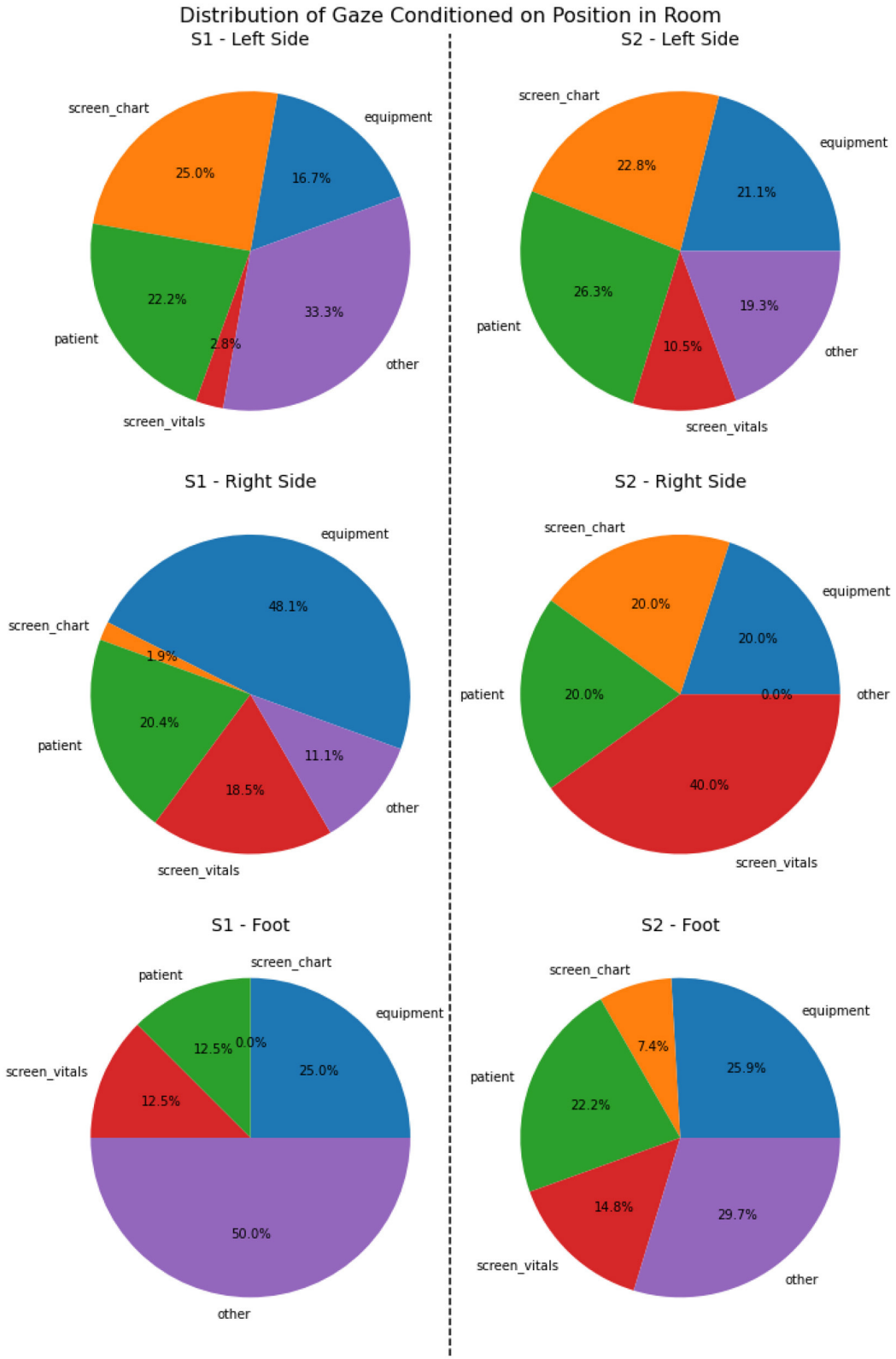

FIGURE 11
Distribution of gaze across five major object categories conditioned on the nurse's position in the room for each scenario.

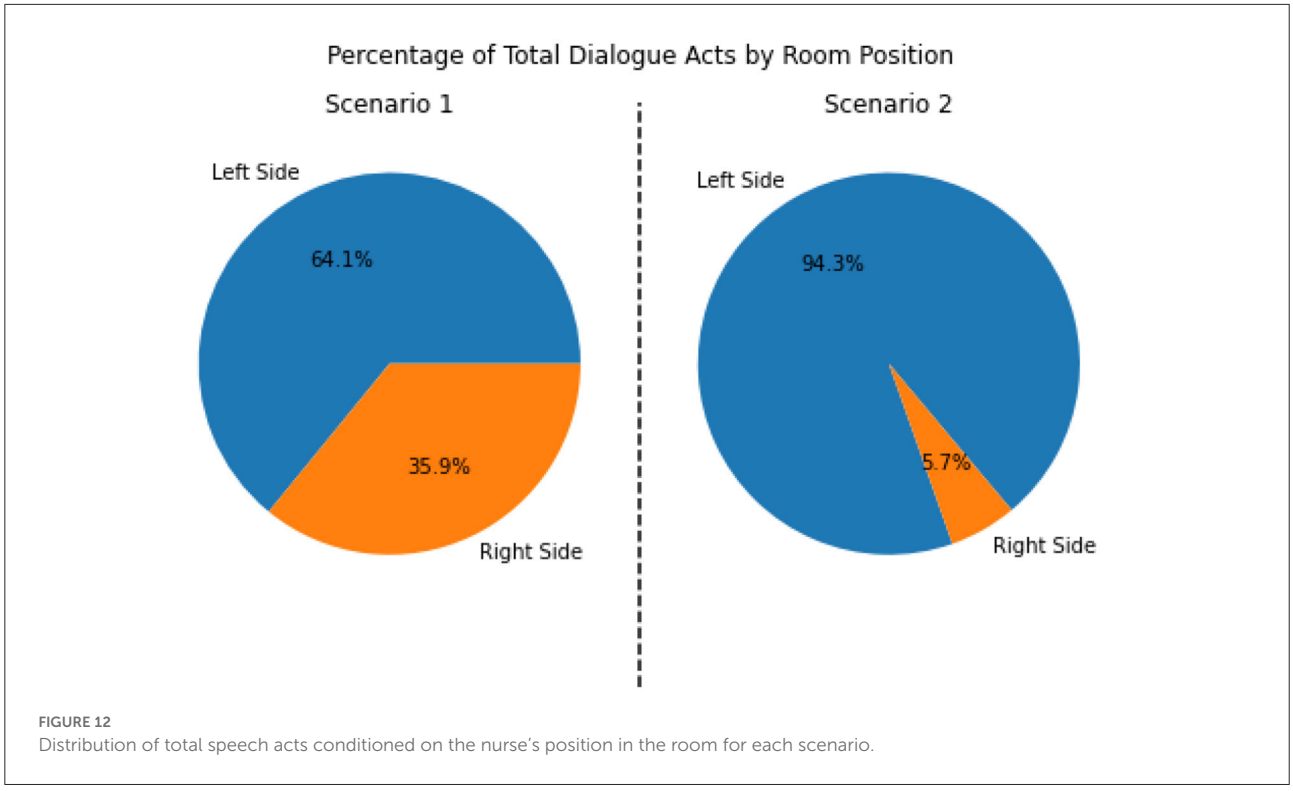

FIGURE 12
Distribution of total speech acts conditioned on the nurse's position in the room for each scenario.

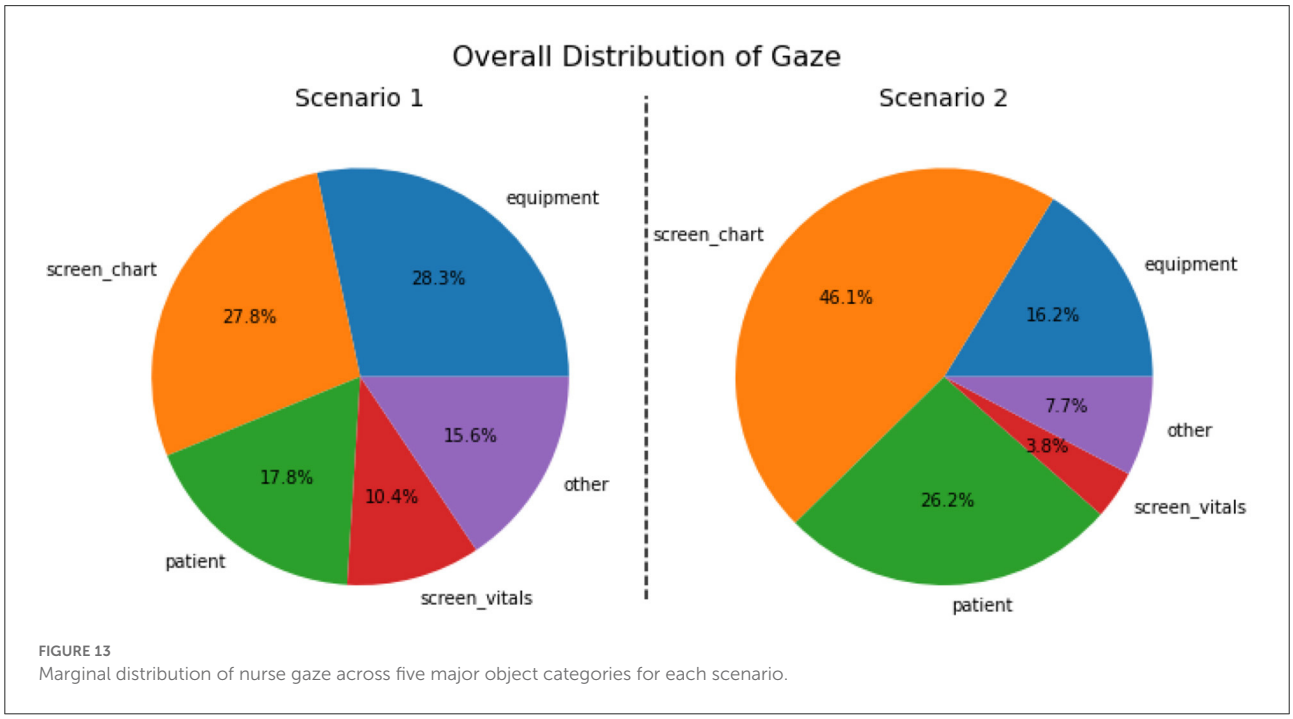

FIGURE 13
Marginal distribution of nurse gaze across five major object categories for each scenario.

when positioned on the left side of the bed. This suggests that the nurses' information gathering done through dialogue with the patient happened primarily when they were on the left side of the bed. Together, this gaze and dialogue data further confirmed the role of each of these spaces in the room; the left side of the bed was primarily used for information gathering and the right side was primarily used for providing clinical procedures.

For the information flow theme, data from the nurse gaze provided significant support for three primary information sources described in the DiCoT analysis: the chart, the patient, and the vitals monitor. Examining Figure 13, which shows the marginal distribution of nurse gaze over the course of the entire scenario. It is clear that the nurses spent most of their time looking at three primary information sources. Over 50% of the total gaze time in both scenarios was spent looking at these three information sources, with 56% for scenario 1 and 76.1% for scenario 2.

The nurses used these three sources to gather, aggregate, and synthesize information which may have been relevant to the patient's diagnosis and treatment. The timeline analysis also supports the information flow theme, demonstrating the transition from information flowing to the nurse to information flowing from the nurse. In both scenarios, the first two timeline segments involve the nurse gathering information. In the first segment, this information came primarily from reading the patient chart and conversation with the patient. In the second segment, the information came primarily from the nurse performing clinical activities.

At this point, the information flow in both scenarios reversed, with the nurses now becoming the information source and the patients and provider becoming the information recipients. Once the nurses had transformed and synthesized the gathered information, they reported their diagnostic inferences, thereby becoming an information source. In both scenarios, the nurse first provided information on her conclusions to the patient, explaining the diagnosis and how they arrived at that conclusion. Then the nurse provided information to the medical provider, first in the form of general patient information over the phone, and then in the form of explaining the diagnosis once the provider arrived in the room.

Moving on to the artifacts and environment theme, the gaze data again clearly supported the use of medical equipment as the primary mediating artifact. As seen in Figure 13, the nurses in both scenarios spent a significant portion of their time with their gaze fixated on the equipment. In scenario 1, the equipment represented the single highest portion of gaze activity at 28.3%. In scenario 2, the nurse looked at the equipment for less time than in scenario 1, but still for a large portion of the total time: 16.2%, which was third overall in terms of the activities conducted. The difference in time here between the two scenarios can be explained by the context of the patient's presenting issue; in scenario 1, the primary cause was primarily linked to the equipment, i.e., the blood pump infusing the wrong blood type. In scenario 2, the primary cause was internal to the patient. The significant portion of time in both scenarios dedicated to medical equipment is evidence of its fundamental role in the distributed cognition analysis of the training scenario.

Beyond these mediating artifacts, the data also supports the use of several artifacts as information hubs, specifically the chart and vitals monitors. As seen in Figure 13, the nurses spend 38.2% and 49.9% of their time, respectively, looking at these two monitors in scenarios 1 and 2. In addition, the timeline analysis suggests that the nurses frequently returned to these information hubs for confirmation and further checking when they were in doubt about their conclusions. For example, we see this behavior in scenario 2 when the nurse looked back at the chart after her physical examination of the patient's left leg did not support her internal diagnostic hypothesis. This shows that the nurses trust the information provided by these artifacts, which support their cognitive reasoning processes in aid of information gathering and diagnostic reasoning.

6. Discussion

Overall, the patterns and distributions derived from our analysis framework clearly demonstrate the effectiveness of our approach in combining qualitative DiCoT analysis with multimodal analytics and the task model to analyze and interpret learner activities and behaviors in the MRMB training simulation. Specifically, this study shows the benefits of our cyclic analysis, with insights generated from both a forward pass of the framework, i.e., using the qualitative analysis to define and structure the quantitative analysis, as well as a backward pass of the framework, i.e., using results of the quantitative analysis to provide more detailed analysis of the learners activities and behaviors than we could generate by pure qualitative analysis, as proposed by the DiCoT framework. The more in-depth information generated by multimodal analysis benefits the two primary stakeholders: (1) learners and instructors through debriefing and after-action reviews, and (2) simulation designers and researchers, who can study the effectiveness of the simulation scripts in promoting effective learning activities. In this section, we discuss the implications of the framework and its resulting insights for both of these groups.

6.1. Implications for learners and instructors

The primary goal of any simulation-based training environment is for the trainees to learn, practice, and develop expertise in skills that transfer to the real task environment. In our nurse case-study, this means that the nurses develop new knowledge and experience that supports both the psychomotor skills and cognitive and metacognitive processes. One of the critical components that mediates this knowledge gain, especially for novice learners, is effective feedback mechanisms during simulation debrief (see Section 2.1). It is the analysis of nurse performance and the generation of relevant feedback

linked to the performance, where our current work is most likely to impact learners and their instructors in constructive ways. By using our analysis framework to generate evaluations of learner behavior, we can present these insights back to learners and instructors during debriefing (also known as after-action reviews) to help promote constructive discussion among the trainees and instructor as part of a larger formative feedback system.

This paper represents an initial step toward analyzing learner performance and behaviors, and then generating formative feedback, and as a result, this case-study analysis was performed *post-hoc*. Therefore, no feedback was generated for learners. However, with continued research, we hope to develop a formative feedback framework with input and support from the instructors to support effective learning of skills and decision making processes. For example, at the beginning of each scenario in our case study, the nurses both start with talking to the patient and reading the patient chart. However, the ways in which these two actions are sequenced differ greatly between the two nurses. In S1, the nurse tended to multi-task combining dialogue with the patient and reading the chart. On the other hand, in S2, the nurse spent a long period of time solely focused on reading the chart without any interaction with the patient, Only after she had reviewed the chart in some detail, did she start talking to the patient in depth. By generating analytics about the nurses' gaze and speech patterns, we can highlight this difference between the nurses and present this feedback as a discussion point during debriefing: Was there a good reason for the difference in approach between the two nurses? Is it not important that the nurse to communicate with the patient sufficiently often so patients do not feel that they are being ignored? Therefore, some level of multi-tasking may be a useful protocol to adopt at this stage of examining the patient and collecting information about their situation. As a next step, we hope to get nursing instructors and experts in as part this discussion. This will help us generate appropriate feedback that will help learners, and also help instructors in setting up constructive discussion among the learners by presenting contrasting cases (Bransford and Schwartz, 1999).

While this is only one simple example, it demonstrates the underlying concept: analytics generated using our activity analysis framework can be presented back to learners and instructors to help promote meaningful discussion, especially around topics that may be otherwise difficult to identify in a single viewing of the scenario. The design of formative feedback that is actionable and important for discussion is a large research questions in itself (Jørnø and Gynther, 2018; Pardo, 2018) and is beyond the scope of this paper; however, analysis framework we present here represents an important first step in this direction for SBT and MRMB training environments.

6.2. Implications for simulation designers and researchers

Because of the cyclic nature of our analysis framework, the insights generated from our analysis and future analytics methods can be used to help refine the qualitative models of the simulation system. This is of particular importance and interest to simulation designers and researchers, as it uncovers new insights to improve our understanding of both the given simulation system and the science of simulation-based training as a whole.

For example, the multimodal data analysis permits the discovery of latent relations between different aspects of the distributed cognition system. In our nursing case-study, this is exemplified through the use of information hubs. The distribution of gaze conditioned on position reveals new insights about the use of information hubs. Initially, the DiCoT analysis revealed the dependency of physical space as a mediator in collection and analysis of the information provided on the two screens as information hubs (i.e., the patient chart and vitals). By combining the physical, artifacts, and information flow segments of DiCoT analysis, we derived how the use of each screen was largely mediated by the nurse's position on the left side of the bed near the patient chart, or on the right side of the bed near the vitals monitor. As described in Section 5.3, we see support for this analysis in the conditional gaze distribution, with fixations on the vitals screen going from 2.8 to 18.5% and 10.5 to 20%, when moving from left to right of the bed in scenarios 1 and 2, respectively.

However, based on this initial DiCoT analysis, we would also expect fixations on the patient chart to have the opposite relationship, decreasing significantly when moving from left to right of the bed. However, in our case studies, the fixations on the patient chart only significantly decreased in S1, moving from 25% on the left to 1.9% on the right, while in S2 the fixations on the patient chart decreased very slightly, with 22.8% on the left and 20% on the right. While it is clear that physical layout mediates the use of these information hubs, the data also suggests an additional latent mediating factor is present. We hypothesize that differences in the simulation scenarios contributed to this, with S2 requiring more references to the patient chart than S1, probably because of the incorrect diagnostic hypothesis the nurse initially made, but there are also other potential explanations, such as differences in the strategies adopted by the two nurses.

This is a simple example of a new insight generated by the quantitative methods that can lead to additional research to refine the qualitative models, but it also demonstrates the overall idea of the cyclic model design. After using the system to analyze learner data, we gain new insights that can be given back to simulation designers and researchers to help formulate new research questions and supporting simulation studies. We can

iteratively update our qualitative understanding of simulation based on learner data, leading to better analysis of the data, and subsequent learner feedback, in the future.

7. Conclusions

In this paper, we presented an analysis of a nurse simulation-based training environment using multimodal learning analytics, cognitive task analysis, and distributed cognition analysis using the DiCoT framework. We show how the analysis of multimodal data from both qualitative and quantitative perspectives can be combined into a common framework for analyzing mixed-reality simulation-based training environments, such as the nursing case study analyzed here. While this work is still in its initial stages, the analysis methods developed and demonstrated here suggest a great potential for combining qualitative distributed cognition analysis with multimodal quantitative analytics in order to generate a more complete understanding of SBT as a whole. The strengths of each method are amplified when used together, and such an integrated approach can help shed new lights on simulation-based training and generate new insights.

However, this work and the framework it presents are not without limitations, and future work is required to address these concerns. One of the major limitations of the presented framework is its lack of guidance on the selection of adequate data sources and design of the associated analysis techniques. Since relevant data sources and analysis techniques differ widely among SBT domains, it is difficult to create a universal guidance on selection and design of these concepts while also keeping the domain-generality of the presented framework. In addition, this study was also limited by the sample size, only analyzing a small case-study of two simulation. This small study size allowed us to focus carefully on the design of the framework and the specific feature of the analysis, but limits the argument for the generalizability of the framework and the analysis results.

Future work will expand our study, both to more data from the nurse training simulation domain, as well as to a variety of other training domains. This expanded work will help to mitigate both of these limitations, as it will allow us to further validate the analysis methods across a wide variety of participants, as well as reveal commonalities among disparate training domains that can be used to generate guiding principles for the selection of adequate data sources and design of the associated analysis techniques. In addition, these further studies will place an emphasis on capturing data related to collaborative and teamwork activities in these environments, helping to further develop the distributed cognition frameworks that ground our data analysis techniques.

To support these expanded studies, future work will also focus on replacing the manual annotation of data used in this study with automated AI and machine learning techniques. Specifically, manual annotation was used in this study for the action, speech, and gaze modalities. For actions, techniques from video activity/action recognition will be applied to automatically extract time segments where the nurse is performing relevant actions (Ghadiyaram et al., 2019; Zhu et al., 2020). For speech, tagging will be automated using pre-trained natural language processing models, such as deep transformer models like Google BERT (Devlin et al., 2018), which have been fine-tuned on our specific domain. In addition, these pre-trained language models will also be applied toward a variety of other downstream NLP tasks, such as event detection and discourse analysis. For gaze, computer vision techniques will be used to automatically match the egocentric video to annotated static camera viewpoints, allowing us automatically determine specific objects (AOIs) that the nurse is looking at (Bettadapura et al., 2015).

Finally, this study and its associated framework was limited in guiding the design of formative learner feedback mechanisms based on the analysis. While Section 6 discussed some of the implications of the framework and its analysis on learning and pedagogy, including the possibility of developing formative learner feedback to support discussion sessions using contrasting cases, the framework itself does not detail guidance for designing learner feedback mechanisms. In addition, for this study specifically, analysis of the case-study data was performed *post-hoc*, so feedback based on the analysis could not be generated in-time for students. Future work, will automate the analysis methods, develop learning analytics to evaluate learner behaviors and actions, and will focus on presenting learners with online feedback designed to support simulation debriefing and after-action reviews. By presenting the results of our analysis to learners and instructors, we can get valuable feedback about the usability of the system and what types of feedback mechanisms might be relevant and important for these stakeholders to see in future iterations of the system.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Vanderbilt University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

GB is the principle investigator of the study and contributed to its initial conceptualization and development of the analysis framework. CV, ED, and CC were responsible for data collection, annotation, and curation. NM maintained the computational and data infrastructure. CV performed the primary data analysis and wrote the initial draft of the manuscript. All authors contributed to model development, interpretation of results, manuscript revision, and approved the submitted version.

Funding

This work represents independent research supported in part by Army Research Laboratory Award W912CG2220001 and NSF Cyberlearning Award 2017000, as well as equipment and funding from the Vanderbilt LIVE initiative.

Acknowledgments

The authors would like to thank Daniel Levin, Madison Lee, and Eric Hall for their instrumental contributions in the design of the study and collection of data. In addition, the authors would like to thank Eric Hall, Jo Ellen Holt, and Mary Ann Jessee for providing their domain expertise during initial planning of the study and during the development of the models used for analysis in this paper. We would also like to thank all of the nursing students who participated in the study. Finally, we would like to thank the reviewers of this paper, who's feedback and guidance strengthened the paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views expressed in this paper do not necessarily reflect the position or policy of the United States Government or the National Science Foundation, and no official endorsement should be inferred.

References

Al-Ghareeb, A. Z., and Cooper, S. J. (2016). Barriers and enablers to the use of high-fidelity patient simulation manikins in nurse education: an integrative review. *Nurse Educ. Today* 36, 281–286. doi: 10.1016/j.nedt.2015.08.005

Bettadapura, V., Essa, I., and Pantofaru, C. (2015). "Egocentric field-of-view localization using first-person point-of-view devices," in *2015 IEEE Winter Conference on Applications of Computer Vision* (Waikoloa, HI: IEEE), 626–633.

Biswas, G., Rajendran, R., Mohammed, N., Goldberg, B. S., Sottilare, R. A., Brawner, K., et al. (2019). Multilevel learner modeling in training environments for complex decision making. *IEEE Trans. Learn. Technol.* 13, 172–185. doi: 10.1109/TLT.2019.2923352

Blandford, A., and Furniss, D. (2006). "Dicot: a methodology for applying distributed cognition to the design of teamworking systems," in *Interactive Systems. Design, Specification, and Verification*, eds S. W. Gilroy and M. D. Harrison (Berlin; Heidelberg: Springer Berlin Heidelberg), 26–38.

Blikstein, P. (2013). "Multimodal learning analytics," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (Leuven), 102–106.

Blikstein, P., and Worsley, M. (2016). Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. *J. Learn. Anal.* 3, 220–238. doi: 10.18608/jla.2016.32.11

Bransford, J. D., and Schwartz, D. L. (1999). Chapter 3: rethinking transfer: a simple proposal with multiple implications. *Rev. Res. Educ.* 24, 61–100. doi: 10.3102/0091732X024001061

Bylinskii, Z., Borkin, M. A., Kim, N. W., Pfister, H., and Oliva, A. (2015). "Eye fixation metrics for large scale evaluation and comparison of information visualizations," in *Workshop on Eye Tracking and Visualization* (Cham: Springer), 235–255.

Clark, A. (1997). *Being There*. Cambridge, MA: MIT Press.

Clark, R. E., and Estes, F. (1996). Cognitive task analysis for training. *Int. J. Educ. Res.* 25, 403–417. doi: 10.1016/S0883-0355(97)81235-9

Cochran, K., Cohn, C., Hutchins, N., Biswas, G., and Hastings, P. (2022). "Improving automated evaluation of formative assessments with text data augmentation," in *AIED* Durham, NC.

Cole, M. (1998). *Cultural Psychology: A Once and Future Discipline*. Cambridge; Massachusetts, MA; London: Harvard University Press.

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., and Brydges, R. (2014). What counts as validity evidence? examples and prevalence in a systematic review of simulation-based assessment. *Adv. Health Sci. Educ.* 19, 233–250. doi: 10.1007/s10459-013-9458-4

Cooper, J., and Taqueti, V. (2008). A brief history of the development of mannequin simulators for clinical education and training. *Postgrad Med. J.* 84, 563–570. doi: 10.1136/qshc.2004.009886

Daniels, K., and Auguste, T. (2013). Moving forward in patient safety: multidisciplinary team training. *Semin. Perinatol.* 37, 146–50. doi: 10.1053/j.semperi.2013.02.004

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv e-prints*, arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805

Di Mitri, D., Schneider, J., Specht, M., and Drachsler, H. (2019). Detecting mistakes in cpr training with multimodal data and neural networks. *Sensors* 19, 3099. doi: 10.3390/s19143099

Feinstein, A. H., and Cannon, H. M. (2002). Constructs of simulation evaluation. *Simulat. Gaming* 33, 425–440. doi: 10.1177/1046878102238606

Fraser, K. L., Ayres, P., and Sweller, J. (2015). Cognitive load theory for the design of medical simulations. *Simulat. Healthcare* 10, 295–307. doi: 10.1097/SIH.0000000000000097

Freedman, D. H. (2010). Why scientific studies are so often wrong: the streetlight effect. *Discover Mag.* 26, 1–4. Available online at: https://www.discovermagazine.com/the-sciences/why-scientific-studies-are-so-often-wrong-the-streetlight-effect

Fu, H., Wu, L., Jian, M., Yang, Y., and Wang, X. (2019). "Mf-sort: simple online and realtime tracking with motion features," in *International Conference on Image and Graphics* (Springer: Beijing), 157–168.

Galliers, J., Wilson, S., and Fone, J. (2007). A method for determining information flow breakdown in clinical systems. *Int. J. Med. Inform.* 76, S113-S121. doi: 10.1016/j.ijmedinf.2006.05.015

Geertz, C. (1973). "The growth of culture and the evolution of mind," in *The Interpretation of Cultures* (New York, NY), 76.

Gegenfurtner, A., Quesada-Pallarès, C., and Knogler, M. (2014). Digital simulation-based training: a meta-analysis. *Br. J. Educ. Technol.* 45, 1097–1114. doi: 10.1111/bjet.12188

Ghadiyaram, D., Tran, D., and Mahajan, D. (2019). "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 12046–12055.

Hazlehurst, B., Gorman, P. N., and McMullen, C. K. (2008). Distributed cognition: an alternative model of cognition for medical informatics. *Int. J. Med. Inform.* 77, 226–234. doi: 10.1016/j.ijmedinf.2007.04.008

Hegland, P. A., Aarlie, H., Strømme, H., and Jamtvedt, G. (2017). Simulation-based training for nurses: systematic review and meta-analysis. *Nurse Educ. Today* 54, 6–20. doi: 10.1016/j.nedt.2017.04.004

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Hum. Interact.* 7, 174–196. doi: 10.1145/353485.353487

Hoppe, H. U. (2017). "Computational methods for the analysis of learning and knowledge building communities," in *Handbook of Learning Analytics* (Beaumont, AB), 23–33.

Hutchins, E. (1991). "The social organization of distributed cognition," in *Perspectives on Socially Shared Cognition* (Washington, DC: American Psychological Association), 283–307.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Hutchins, E. (2000). "Distributed cognition," in *International Encyclopedia of the Social and Behavioral Sciences* (Amsterdam: Elsevier Science), 138.

Hutchins, E. (2006). The distributed cognition perspective on human interaction. *Roots Hum. Soc.* 1, 375. doi: 10.4324/9781003135517-19

Johnson, M. P., Hickey, K. T., Scopa-Goldman, J., Andrews, T., Boerem, P., Covec, M., et al. (2014). Manikin versus web-based simulation for advanced practice nursing students. *Clin. Simulat. Nurs.* 10, e317-e323. doi: 10.1016/j.ecns.2014.02.004

Jørnø, R. L., and Gynther, K. (2018). What constitutes an 'actionable insight' in learning analytics? *J. Learn. Anal.* 5, 198–221. doi: 10.18608/jla.2018.53.1310.18608/jla.2018.53.13

Kang, S. J., and Min, H. Y. (2019). Psychological safety in nursing simulation. *Nurse Educ.* 44, E6-E9. doi: 10.1097/NNE.0000000000000571

Kaplan, A. D., Cruit, J., Endsley, M., Beers, S. M., Sawyer, B. D., and Hancock, P. (2021). The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: a meta-analysis. *Hum. Factors* 63, 706–726. doi: 10.1177/0018720820904229

Kim, J. W., Sottilare, R. A., Brawner, K., and Flowers, T. (2018). "Integrating sensors and exploiting sensor data with gift for improved learning analytics," in *Sixth Annual GIFT Users Symposyum* Orlando, FL.

Kunst, E. L., Mitchell, M., and Johnston, A. N. (2016). Manikin simulation in mental health nursing education: an integrative review. *Clin. Simulat. Nurs.* 12, 484–495. doi: 10.1016/j.ecns.2016.07.010

Laerdal Medical (2022a). *LLEAP - Laerdal Learning Application* Stavanger: Laerdal Medical.

Laerdal Medical (2022b). *SimMan 3G Advanced Patient Simulator* Stavanger: Laerdal Medical.

Liu, B., Zhao, Q.-C., Ren, Y.-Y., Wang, Q.-J., and Zheng, X.-L. (2018). An elaborate algorithm for automatic processing of eye movement data and identifying fixations in eye-tracking experiments. *Adv. Mech. Eng.* 10, 1687814018773678. doi: 10.1177/1687814018773678

López, M. X., Strada, F., Bottino, A., and Fabricatore, C. (2021). "Using multimodal learning analytics to explore collaboration in a sustainability co-located tabletop game," in *European Conference on Games Based Learning* (Brighton: Academic Conferences International Limited), 482-XXI.

Maran, N. J., and Glavin, R. J. (2003). Low-to high-fidelity simulation-a continuum of medical education? *Med. Educ.* 37, 22–28. doi: 10.1046/j.1365-2923.37.s1.9.x

Martinez-Maldonado, R., Echeverria, V., Fernandez Nieto, G., and Buckingham Shum, S. (2020a). "From data to insights: a layered storytelling approach for multimodal learning analytics," in *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems* (Honolulu, HI), 1–15.

Martinez-Maldonado, R., Elliott, D., Axisa, C., Power, T., Echeverria, V., and Buckingham Shum, S. (2020b). "Designing translucent learning analytics with teachers: an elicitation process," in *Interactive Learning Environments*, 1–15. doi: 10.1080/10494820.2019.1710541

Meakim, C., Boese, T., Decker, S., Franklin, A., Gloe, D., Lioce, L., et al. (2013). Standards of best practice: simulation standard i: Terminology. *Clin. Simulat. Nurs.* 9, S3-S11. doi: 10.1016/j.ecns.2013.04.001

Militello, L. G., and Hutton, R. J. (1998). Applied cognitive task analysis (acta): a practitioner's toolkit for understanding cognitive task demands. *Ergonomics* 41, 1618–1641. doi: 10.1080/001401398186108

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., and Del Maestro, R. F. (2020). The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE* 15, e0229596. doi: 10.1371/journal.pone.0229596

Ochoa, X., Lang, A. C., and Siemens, G. (2017). Multimodal learning analytics. *Handbook Learn. Anal.* 1, 129–141. doi: 10.18608/hla17.011

Olsen, A. (2012). *The tobii ivt fixation filter algorithm*. Los Altos, CA: Technical report, Tobii Pro.

Otter.ai (2022). *Otter.ai-Voice Meeting Notes and Real-time Transcription*. Los Altos, CA.

Pardo, A. (2018). A feedback model for data-rich learning experiences. *Assess. Evaluat. Higher Educ.* 43, 428–438. doi: 10.1080/02602938.2017.1356905

Park, J. E., and Kim, J.-H. (2021). Nursing students experiences of psychological safety in simulation education: a qualitative study. *Nurse Educ. Pract.* 55, 103163. doi: 10.1016/j.nepr.2021.103163

Pimmer, C., Pachler, N., and Genewein, U. (2013). Reframing clinical workplace learning using the theory of distributed cognition. *Acad. Med.* 88, 1239–1245. doi: 10.1097/ACM.0b013e31829eec0a

Ravert, P. (2002). An integrative review of computer-based simulation in the education process. *Comput. Inform. Nurs.* 20, 203–208. doi: 10.1097/00024665-200209000-00013

Rokhsaritalemi, S., Sadeghi-Niaraki, A., and Choi, S.-M. (2020). A review on mixed reality: current trends, challenges and prospects. *Appl. Sci.* 10, 636. doi: 10.3390/app10020636

Rosen, M. A., Salas, E., Wilson, K. A., King, H. B., Salisbury, M., Augenstein, J. S., et al. (2008). Measuring team performance in simulation-based training: adopting best practices for healthcare. *Simulat. Healthcare* 3, 33–41. doi: 10.1097/SIH.0b013e3181626276

Rybing, J. (2018). *Studying Simulations With Distributed Cognition, Vol. 1913*. Linköping: Linköping University Electronic Press.

Rybing, J., Nilsson, H., Jonson, C.-O., and Bang, M. (2016). Studying distributed cognition of simulation-based team training with dicot. *Ergonomics* 59, 423–434. doi: 10.1080/00140139.2015.1074290

Rybing, J., Prytz, E., Hornwall, J., Nilsson, H., Jonson, C.-O., and Bang, M. (2017). Designing a digital medical management training simulator using distributed cognition theory. *Simulat. Gaming* 48, 131–152. doi: 10.1177/1046878116676511

Sawyer, T. L., and Deering, S. (2013). Adaptation of the us army's after-action review for simulation debriefing in healthcare. *Simulat. Healthcare* 8, 388–397. doi: 10.1097/SIH.0b013e31829ac85c

Schraagen, J. M., Chipman, S. F., and Shalin, V. L. (2000). *Cognitive Task Analysis*. New York, NY: Psychology Press.

Stanton, N. A. (2014). Representing distributed cognition in complex systems: how a submarine returns to periscope depth. *Ergonomics* 57, 403–418. doi: 10.1080/00140139.2013.772244

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). "brat: a web-based tool for NLP-assisted text annotation," in *Proceedings of the Demonstrations Session at EACL 2012* (Avignon: Association for Computational Linguistics), 102–107.

Sun, Z., Chen, J., Chao, L., Ruan, W., and Mukherjee, M. (2020). A survey of multiple pedestrian tracking based on tracking-by-detection framework. *IEEE Trans. Circ. Syst. Video Technol.* 31, 1819–1833. doi: 10.1109/TCSVT.2020.3009717

Tobii Pro (2012). *Determining the Tobii I-VT Fixation Filter's Default Values*. Danderyd Municipality: Tobii Technology

Tobii Pro (2022). *Tobii Pro Glasses 3*. Danderyd Municipality: Tobii Technology.

Vatral, C., Biswas, G., and Goldberg, B. S. (2022). "Multimodal learning analytics using hierarchical models for analyzing team performance," in *Proceedings of the 15th International Conference on Computer Supported Collaborative Learning* (Hiroshima: International Society of the Learning Sciences).

Vatral, C., Mohammed, N., Biswas, G., and Goldberg, B. S. (2021). "Gift external assessment engine for analyzing individual and team performance for dismounted battle drills," in *Proceedings of the 9th Annual Generalized Intelligent Framework for Tutoring User Symposium* (Orlando, FL: US Army Combat Capabilities Command Center), 109–127.

Wojke, N., Bewley, A., and Paulus, D. (2017). "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)* (Beijing: IEEE), 3645–3649.

Worsley, M., and Martinez-Maldonado, R. (2018). "Multimodal learning analytics' past, present, and potential futures," in *CrossMMLA@ LAK* (Sydney), 1–16.

Wright, P. C., Fields, R. E., and Harrison, M. D. (2000). Analyzing human-computer interaction as distributed cognition: the resources model. *Hum. Comput. Interact.* 15, 1–41. doi: 10.1207/S15327051HCI1501_01

Zachary, W. W., Ryder, J. M., and Hicinbothom, J. H. (2000). "Building cognitive task analyses and models of a decision-making team in a complex real-time environment," in *Cognitive Task Analysis* (New York, NY), 365–384.

Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., et al. (2020). A comprehensive study of deep video action recognition. *arXiv preprint* 1–30. doi: 10.48550/arXiv.2012.06567

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 08 August 2022
DOI 10.3389/frai.2022.891630

OPEN ACCESS

EDITED BY
Paul Seitlinger,
University of Vienna, Austria

REVIEWED BY
Malinka Ivanova,
Technical University of Sofia, Bulgaria
Matthew Klenk,
Toyota Research Institute,
United States

*CORRESPONDENCE
Bert Bredeweg
b.bredeweg@uva.nl

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 08 March 2022
ACCEPTED 21 July 2022
PUBLISHED 08 August 2022

CITATION
Bredeweg B and Kragten M (2022) Requirements and challenges for hybrid intelligence: A case-study in education. *Front. Artif. Intell.* 5:891630. doi: 10.3389/frai.2022.891630

Requirements and challenges for hybrid intelligence: A case-study in education

Bert Bredeweg[1,2]* and Marco Kragten[1]

[1]Faculty of Education, Amsterdam University of Applied Sciences, Amsterdam, Netherlands, [2]Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, Netherlands

The potential for Artificial Intelligence is widely proclaimed. Yet, in everyday educational settings the use of this technology is limited. Particularly, if we consider smart systems that actually interact with learners in a knowledgeable way and as such support the learning process. It illustrates the fact that teaching professionally is a complex challenge that is beyond the capabilities of current autonomous robots. On the other hand, dedicated forms of Artificial Intelligence can be very good at certain things. For example, computers are excellent chess players and automated route planners easily outperform humans. To deploy this potential, experts argue for a hybrid approach in which humans and smart systems collaboratively accomplish goals. How to realize this for education? What does it entail in practice? In this contribution, we investigate the idea of a hybrid approach in secondary education. As a case-study, we focus on learners acquiring systems thinking skills and our recently for this purpose developed pedagogical approach. Particularly, we discuss the kind of Artificial Intelligence that is needed in this situation, as well as which tasks the software can perform well and which tasks are better, or necessarily, left with the teacher.

KEYWORDS

Qualitative Reasoning, science education, systems thinking with qualitative representations, real-world application problems, hybrid human-AI systems

Introduction

The expected added value of Artificial Intelligence was already high at its inception (McCarthy et al., 1955). Meanwhile, impressive results have been obtained, but these solutions are typically highly specialized (e.g., Silver et al., 2016). The realization of Artificial General Intelligence (AGI) or strong Artificial Intelligence (e.g., Kurzweil, 2005) has not yet happened, and it may take a long time for it to happen (Marcus and Davis, 2019). Instead of aiming for AGI, the idea of Hybrid Intelligence is being proposed (Akata et al., 2020). Hybrid Intelligence combines human intelligence with machine intelligence, with the goal of augmenting human capabilities as opposed to replacing them, while simultaneously harvesting the potential of smart machines.

In the area of Intelligent Tutoring Systems, which was traditionally highly focused on automating tutoring to the max (Wenger, 1987), such alternative hybrid approaches are also discussed. Chou et al. (2011) report a study in which two virtual

teaching assistants successfully aid the teacher. One assistant focuses on evaluating student's answers and the other on generating hints. Baker (2016) argues that successful automated tutoring systems do not show general (teaching) intelligence, but rather excel in a specific capability. As such, they emphasize the use of educational data mining to support human-decision-making. Another example of a hybrid approach is the work of Paiva and Bittencourt (2020) who implemented an authoring tool that deals with educational data from an online course to support instructors in making pedagogical decisions. Holstein et al. (2019) report on a study that investigates students and teachers needs with regard to human vs. Artificial Intelligence instruction help-signaling and help-giving. They found that teachers desire greater real-time support from the automated tutors, and that students emphasize their need for help-signaling without losing face to peers. Holstein et al. (2020) present a framework consisting of a set of dimensions that describe how hybrid teacher/AI adaptivity can augment performance and enhance co-learning on instructional goals, relevant information, instructional actions and decisions.

The examples show that the concept of Hybrid Intelligence in education is being discovered. Additional studies and real-life applications may help to further understand and develop this approach. In this contribution, we report on a case study that uses smart tutoring software in secondary education. While Intelligent Tutoring examples often focus on problem solving, we focus on learning by creating qualitative representations. Learners learn systems thinking by creating a diagram that captures a causal understanding of how a system works. Different from typical problem assignments, in which case the solution amounts to a specific answer such as a number after having performed the required calculations, learners create and deliver a structure consisting of a set of ingredients and relationships among these (Spitz et al., 2021a).

The organization of this paper is as follows. Section 'What makes a system Artificial Intelligent?' briefly reviews the field of Artificial Intelligence research in order to define what we mean when we refer to an Artificial Intelligence system. Next, we move to the case study in which learners in secondary education acquire systems thinking skills and the hybrid teacher-software arrangement to support that. Section 'The case-study: An automated intelligent systems thinker in secondary education' describes our recently developed intelligent tutoring system and the accompanying pedagogical approach that supports learners in creating their cause-and-effect diagrams. Section 'Teacher's role' discusses the role of the teacher and how it complements and intertwines with the actions of the tutoring system. Section 'Conclusions and discussion' concludes this contribution and Section Future work highlights directions for future research.

What makes a system Artificial Intelligent?

It remains intriguing to observe computers solve problems that up to then only people could solve well. Even more when the computer solves versions of those problems that it has not been given explicitly before. On the other hand, the ubiquitous pocket calculator is generally not discussed as an example of smart software, even though it outperforms most humans when it comes to doing mathematics. What is it that characterizes Artificial Intelligence since it came into existence in the 60s?

The reoccurring trinity

Let us start with a short historical perspective. One inspiration for Artificial Intelligence originates from Psychology. When the cognitivist paradigm (Lindsay and Norman, 1977) succeeded the behaviorist paradigm (Skinner, 1974), computer programs became fashionable as cognitive or mental models (Gentner and Stevens, 1983). A requirement for developing cognitive models is to make the solution generic. Instead of being able to solve one specific example, a viable solution is capable of solving all possible instances of the problem.

Over the years, many ingenious algorithms have been invented (Bratko, 2012). Additionally, the importance of adequate representations became recognized, both the formal language (the knowledge representation language) and the representation of substantive knowledge in it (the knowledge base). The endeavor grew into automating miscellaneous kinds of human expertise such as the expertise of chess players, physicians, designers, etc. (Schreiber et al., 1993). A noteworthy milestone was reached in 1987 when IBM's Deep Blue II program successfully defeated Kasparov, the then reigning chess grandmaster. Notice that, improved hardware was also a key enabler for this milestone (BNVKI, 2021).

The wealth of ideas and approaches is enormous (van Harmelen et al., 2008) and the area is still advancing (Moschoyiannis et al., 2021). This also holds for the work on cognitive systems (Nirenburg, 2017). If, in hindsight, we consider the overarching research agenda, it becomes apparent that Artificial Intelligence works on three key questions (see also Figure 1):

- How to represent? The focus here is on the development of (semi-)formal languages, typically referred to as a *Knowledge Representation Language (KRL)*. Essentially a set of interrelated concept types that together conform

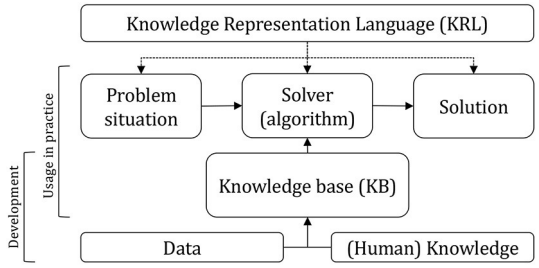

FIGURE 1
AI system characteristics. When developed, knowledge or data is obtained from humans and/or data resources, processed and stored into a Knowledge Base (KB). The KB adheres to a Knowledge Representation Language (KLR) which is also the basis for the development of the solver. When used in practice the solver receives a case (problem situation), and deploys the KB to develop a solution.

to a certain semantics and that can be used to store (or represent) pieces of information.[1]

- What to represent? The process of selecting a certain amount of knowledge (or information), untangling it into elementary parts in accordance with the KRL and storing it. The process can be executed by humans, but also (partly) automated using software. The result is typically referred to as the *Knowledge Base (KB)*.
- How to reason? The development of *solvers* or *algorithms*, often tailored toward the specifics of the KRL, and their deployment to solve problems. Concerning the latter, the algorithm obtains or receives information about an actual case or *problem situation* and is able to draw conclusions (*solution*) by relating this input to the KB and making the appropriate inferences.

Depending on the actual implementation the appearance and use of an Artificial Intelligence system can be highly different. For instance, an automated agent continuously regulating some system as opposed to a classifier that each runtime produces a particular output.

Neural networks are also among the early ideas researched within the context of Artificial Intelligence (McCullogh and Pitts, 1943; Rosenblatt, 1958). With the arrival of abundant data and significant faster hardware, neural networks are now also well developed (LeCun et al., 2015; Goodfellow et al., 2016). They received much attention since the computer won the game of Go (Silver et al., 2016). Although the proclaimed potential is also critically reviewed (Marcus and Davis, 2019). Neural networks also adhere to the above described trinity: (i) there is a representation language consisting of interconnected units (referred to as neurons, layers, etc.), (ii) there is a body of information stored using this representation (typically, build from a huge set of examples), and (iii) there is an algorithm that reasons about specific cases using this stored information. The creation of the stored information, the "knowledge base", can be automated in the case of isolated, formal contexts (Silver et al., 2017). However, for real applications the organization of data (data wrangling) is a complex and time-consuming task, typically performed by human experts (e.g., Kuhn and Johnson, 2019).

Truly intelligent?

As discussed above, systems referred as to Artificial Intelligence concern three intertwined components: the representation language, the stored content, and the reasoning. When these components are well established, an artificial system can be deployed in the real-world situation for which it was developed, where it will behave according to its capacity.

A number of concerns associated with intelligent behavior are often brought up when (thinking about) using Artificial Intelligence in practice (e.g., Marcus and Davis, 2019; Aicardi et al., 2020):

- Specialization. It is generally known that Artificial Intelligence systems are highly specialized (or limited, if one prefers) and only work well for the specifics they were developed for. A system aiding physicians in finding deviating spots in x-rays, will do exactly that, and nothing else.
- Reliability and trustworthiness. Exactly when will the system fail? Is it capable of handling all the potential cases correctly? Can the software be trusted? Will it behave ethically? Notice that, the software itself typically has no clue regarding its own competence and actions, nor its limitations.
- Transparency and explanation. Can the software explain its reasoning? Explain how it came to a certain result? Moreover, can the software argue why a result or conclusion is correct or viable? In fact, the dichotomy between effective reasoning vs. insightful explanations thereof, is a long standing challenge in Artificial Intelligence.

Does having these limitations make an Artificial Intelligence less smart? Do humans not have similar limitations? Why do we want to regard human-made computer software as intelligent in the first place? These are difficult questions to answer. In fact, the answers depend on the perspective taken. The categorization of Artificial Intelligence as put forward by Russell and Peter Norvig (2020) is helpful in this respect. Instead of emphasizing

1 The difference between data, information and knowledge is subtle. Here we use these terms interchangeably and only make explicit distinctions when needed.

TABLE 1 Which intelligence do Artificial Intelligence systems need?

	Pocket calculator	Cognitive system	Medical diagnosis	Education
Thinking humanly		?		
Acting humanly		x	?	?
Thinking rationally	x		x	x
Acting rationally			?	?

a particular technology or a characteristic of intelligence, their focus is on the reference to which the solution is compared. Consequently, multiple kinds of Artificial Intelligence research and applications exist:

- Thinking humanly: The cognitive modeling approach.
- Acting humanly: The Turing test approach.
- Thinking rationally: The "laws of thought" approach.
- Acting rationally: The rational agent approach.

Table 1 shows examples to further illustrate this framework. Consider the pocket calculator mentioned earlier. We can argue that it "thinks" fully rational, following the rules of mathematics. As such, we should acknowledge that it implements a form of intelligence, even though it is not considered a typical Artificial Intelligence system (among others, it misses the knowledge base component discussed above). For any cognitive system (a system oriented toward human skills and capabilities) it should at least act humanly (e.g., pass the Turing test) and dependent on the (research) goal possible also think humanly. For an application of Artificial Intelligence supporting a physician in doing medical diagnosis we would definitely require a fully rational thinking machine. If the robot is also expected to interact with patients, maybe it should also have features of acting humanly. If it is also expected to be a pro-active and caretaking system, an autonomous robot that acts rationally is probably wanted. Similar arguments hold for intelligent applications in education. Foremost, it should be a rational thinking machine that is capable of handling the subject matter in interaction with learners. If it is also expected to be pro-active in the class, or even take a leading role, an autonomous rationally acting robot will be needed. Should it also act humanly? Maybe, but such behavior may also hamper optimal teaching behavior. After all, typical human behavior, even that of experts, may not always be the best solution in a challenging situation (Holstein et al., 2019).

The case-study: An automated intelligent systems thinker in secondary education

Let us now discuss an Artificial Intelligence system in education. As a case-study, we focus on learners acquiring systems thinking skills. Systems thinking is an important skill for humans to master (e.g., NGSS, 2013), but difficult to learn (e.g., Sweeney and Sterman, 2007). We develop and investigate a new pedagogical approach to having learners in secondary education acquire this skill using qualitative representations (https://denker.nu/). The approach covers K8-12 and is linked to the curriculum in the subjects of biology, physics, geography and economics. Table 2 gives an overview of the main tasks involved and the distribution among the participants (including the intelligent software, AI-App).

For education the goal is to create smart people and the Artificial Intelligence is used as a tool to enable that. There are at least two reasons why qualitative representations form an interesting set of intelligent tools for education. Firstly, as with any representation, when used by people representations strongly steer the development of knowledge and insights (Davis et al., 1993). As such, having learners construct representations is a valuable pedagogical instrument for implementing active learning (Prain and Tytler, 2012). Secondly, the Qualitative Reasoning community particularly focused on explicating the implicit knowledge considered essential for reasoning about the behavior of (physical) systems. This resulted in an explicit vocabulary underpinning automated reasoning. In fact, the community developed an explicit ontology (Liem, 2013) for (automated) systems thinking. Modern educators emphasize the importance and challenge of supporting learners in lower and upper secondary education in acquiring systems thinking skills (Jacobson and Wilensky, 2006; Ben-Zvi-Assaraf and Orion, 2010; Curriculum.nu, 2021). Qualitative representations can be deployed for this purpose. Their suitability is even more profound because of the accompanying automated reasoners, which makes them outstanding candidates for intelligent interactive tools for learning systems thinking.

Our approach is based on a classroom situation with on average 30 learners and a teacher. Additionally, it includes an intelligent software for creating qualitative representations and a workbook to guide learners and teachers during this process. The role the software, particularly in relation to the learner, is described below. Section 'Teacher's role' describes the role of the teacher.

TABLE 2 Summarizing overview of tasks, including task description, the executing agent, the resources used to accomplish the task, the output the task delivers, and the beneficiary who uses the output.

Task description	Agent	Resources	Output	Beneficiary
Creating a knowledge-base for a curriculum topic	Teacher	Curriculum	KB-norm	AI-App
Creating a workbook	Teacher	Curriculum; KB-norm; AI-App	Workbook	Learner
Engaging in a dialogue to address a knowledge deficiency	Teacher	Learner request; KB-learner; AI-App; Expert knowledge	Advanced explanation	Learner
Managing the classroom and engaging learners	Teacher	Class behavior and history; Learner characteristics	Effective learning environment	Learner
Learning by creating a representation	Learner	Workbook; AI-App	KB-learner	AI-App; Teacher
Calling AI-App to compute system behavior for KB-learner	Learner	KB-learner; AI-App	Inferred system behavior	Learner
Asking for help on a knowledge deficiency	Learner	Workbook; KB-learner; Inferred system behavior	Learner request	Teacher
Finding deviating ingredients in KB-learner (norm-based cueing)	AI-App	KB-norm; KB-learner	Discrepancies highlighted	Learner
Typing deviating ingredients in KB-learner (norm-based advice)	AI-App	KB-norm; KB-learner; Error type recognizer	Discrepancies error-typed	Learner
Identifying and summarizing correct, incorrect and missing ingredients in KB-learner	AI-App	KB-norm; KB-learner; Discrepancies	Progress bar	Learner
Finding discrepancies in initial settings when calling AI-App	AI-App	KB-learner; Initial settings requirements	Advice on problem situation	Learner
Finding feedback-loop in KB-learner	AI-App	KB-learner; Feedback-loop recognizer	Feedback-loops highlighted	Learner
Describing and predicting learners' learning behavior	AI-App	KB-learner; Action-Log; Automated statistics	Overview of learners' learning behavior	Teacher

KB refers to the representations (knowledge-base) created by the teacher (KB-norm) and by the learner (KB-learner). AI-App refers to the set of algorithms implemented in the AI software.

Knowledge representation language and reasoning (solver)

The software implements an automated intelligent systems thinker (Bredeweg et al., 2009). It builds on research from Artificial Intelligence known as Qualitative Reasoning (Weld and de Kleer, 1990; Forbus, 2018). The KRL consist of ∼15 concepts to describe dynamic systems, including notions such as entity, quantity, value, change, causality, in/equality, etc. The KRL is logic-based and does not use any numerical information. The main reasoning task of the solver is prediction of system's behavior, which includes a whole range of specific algorithms implementing subtasks, such as influence resolution, inequality reasoning, reasoning with assumptions, reasoning with inheritance, etc.

To be an effective tool for learning, it is important to acknowledge that systems thinking is a complex skill. It requires an approach in which the skill is gradually build up. From that perspective it is relevant to realize that the subject matter currently taught in secondary education is also complex and learned stepwise across multiple years as specified by the curricula. In accordance with these constraints, the automated systems thinker is organized such that it is able to work at distinct levels of complexity. There are five levels in total, roughly corresponding to the complexity needed in grade 8–12 (Bredeweg et l., 2010).

Knowledge base

As discussed in Section 'The reoccurring trinity', a well-developed KB is a typical component of Artificial Intelligence systems. In other work, we developed such KBs (e.g., Bredeweg and Salles, 2009). However, dealing with education brings different requirements. An important insight from working with teachers has been that they have specific constraints regarding what their learners need to learn, typically following the details as specified in the curricula. For a smart tool to successfully collaborate with teachers in educating learners, this tool should be adjustable to these requirements. However, covering all the

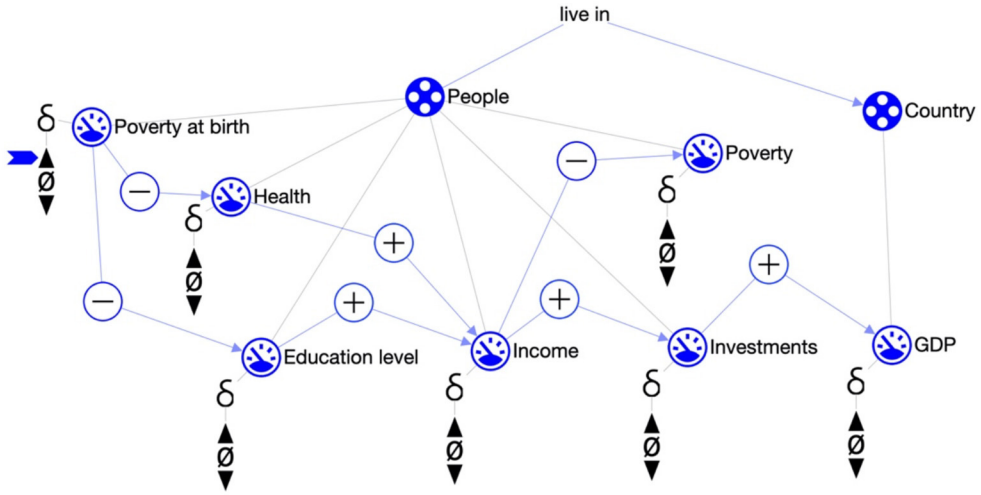

FIGURE 2
Poverty (Spitz et al., 2021b). *Entities* are *People* and *Country*, whose *relationship* is *live in*. There are seven *quantities*, such as *health* and *income*. Quantities are related by positive (+) and negative (−) *causal* relationships. The direction of change is denoted by the δ. In this case, it is only specified for *poverty at birth*, which is set to be increasing.

material present in the text books for all the subjects, and somehow managing that a specific part of that material gets in focus during a particular lesson, is simply not a realistic goal. Hence, we decided to take a different approach and develop small KBs. Each one is dedicated to a specific lesson and accompanying learning goals, and developed in collaborating with and as required by the teachers participating in the project (see also Section 'Subject matter selection and preparation'). A small example is show in Figure 2 (Poverty, developed for geography in grade 8). See for more examples and details Kragten et al. (2021) and Spitz et al. (2021b).

Supporting the learners in acquiring system thinking skills

The tool described in Sections 'Knowledge representation language and reasoning (solver)' and 'Knowledge base' (the automated systems thinker) can be given to learners to support them in their learning process. Essentially, learners learn by creating their own small "knowledge base", mimicking the KB created by the teacher. Both the KRL and the solver are instruments that support the learner in doing so.

Notice, that the knowledge representation is shown to the learner as an **interactive diagram** (a kind of knowledge graph, similar as shown in Figure 2). After its initial design and implementation (e.g., Bouwer and Bredeweg, 2010) this diagrammatic representation has been further developed. Currently, it depicts all the ingredients present in the KRL, and also in the reasoning output, and it enables the learners to interact with these. As such, this graphical format is an important asset, because it hides low-level details and enables learners to work at the "content level".

Having the graphical user interface, and the rest of the underlying tooling [as discussed in Section 'Knowledge representation language and reasoning (solver)'], learners can now independently work on assignments and successfully complete these. However, learners may make mistakes and potentially learn incorrect details or get stuck in executing the assignment. Hence, the teacher has to be alert, monitor and assess the progress of the learners, and intervene where deemed necessary. Although maybe doable in small classes, it does make the teaching laborious for the teacher. To alleviate this burden, we have developed automated reasoners to further support the learner by providing just in-time feedback and to stimulate learners' self-reliance.

Norm-based cueing and advice

The KB discussed in Section 'Knowledge base', which is created together with the teacher, can be used as a norm. Our current implementation compares the learner-created "knowledge-base" (KB-learner) with the KB created by the teacher (KB-norm). After each manipulation executed by the learner in the canvas a new mapping is made using a Monte-Carlo-based heuristic approach. The engine runs for at most 5 s and then returns the best mapping. Next, for each discrepancy the support provides two options for

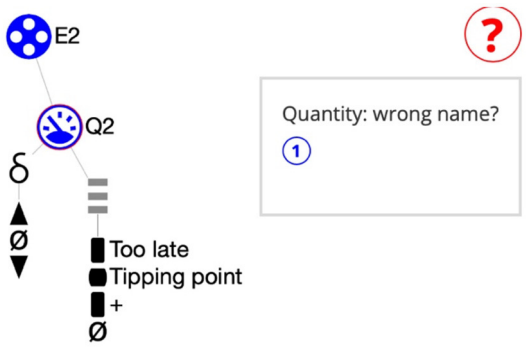

FIGURE 3
Cueing and Advice (Spitz et al., 2021a). While creating the details a quantity is wrongly named. *Cueing* highlights the erroneous ingredient (here: Q2). *Advice* suggests an error. Here: Quantity: wrong name.

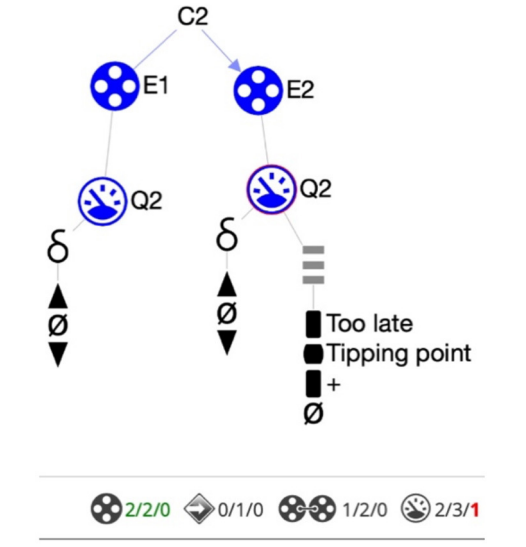

FIGURE 4
Progress bar (partly shown). The status is shown for each ingredient type at the bottom of the canvas. For instance, "Quantities 2/3/1" tells the learner that 2 quantities have been created, 3 need to be created in total and that 1 is currently incorrect (shown in red). When all ingredients have been created correctly the numbers become green, as for entities here.

feedback. **Cueing**: a small red circle is placed around each deviating ingredient (Q2 in Figure 3) and a red question mark appears on the right-hand side in the canvas. **Advice**: when clicking on the question mark, a message-box appears showing a sentence for each deviation (in Figure 3: Quantity: Q2: wrong name?). Note that, the algorithm works domain independent, yet learners get subject specific information. For instance, whether they assign the correct quantities to each of the entities.

Progress bar

Next to being informed about errors, it is also helpful for learners to get information on the degree to which they have accomplished the goals. This will support them in knowing what still needs to be done and when the goal is reached, and may also be relevant to stimulate metacognitive reflection. Our current approach implements the idea of a progress bar (Figure 4). For each ingredient *type* present in the KB-norm the bar shows (*i*) how many instances of that ingredient need to be created, (*ii*) how many at any given moment have been created, (*iii*) how many of those created are incorrect, and (*iv*) when all the details for that ingredient type are addressed (by changing font color to green). Further research is needed to find out whether this support is helpful and sufficient, without giving away too much.

Two further supports are available. The **scenario advisor** inspects the status of the problem situation when presented to the solver by the learner. If errors occur, the advisor will discover these and notify the learner. Examples are, missing initial values for quantities at the start of a causal chain and superfluous values defined for any intermediate quantity, including incorrect values that block possible outcomes from being inferred. The **feedback-loop identifier** highlights loops after the reasoning has delivered the simulation results. Two versions exist, positive feedback (change is reinforced) and negative feedback (change is reduced). The highlights are intended to help learners observe important features in the simulated system's behavior. They can also be used for coaching and further instruction.

Analytics—Supporting the teacher

A learning analytics module has also been developed but not used in practice yet. The aim is to provide the teacher descriptive and predictive overviews, based on the progress learners make measured by the number of correct and incorrect ingredients, number of support agents calls, construction speed, etc.

Teacher's role

Being an effective teacher is a serious challenge (Rosenshine, 2012). The key task is to create an environment that enables a group of, often diverse, learners to successfully develop their knowledge and skills. The size and complexity of this task is currently far beyond the capabilities of any automated agent based on Artificial Intelligence. However, a hybrid approach can be very effective when carefully planned and arranged, especially in specific situations. As discussed above, here we focus on lessons in systems thinking, where on average learners complete a lesson series about a specific topic in ∼2 h. Which tasks does the teacher have, when using an intelligent software in this context?

FIGURE 5
Fragment of the poverty example (Spitz et al., 2021b). The initial situation (LHS) reads as follows: Entity people has two quantities, Income and Poverty. Income is set to increase and has a negative influence on Poverty. Consequently, when simulated (RHS), poverty decreases as can be inferred from the representation.

Subject matter selection and preparation

The subject matter of the lesson must be selected and prepared by the teacher. It involves (*i*) selecting learning goals for content knowledge and system thinking and thereby scoping the learning experience as a whole, (*ii*) creating a qualitative model to serve as the norm for the intelligent agent, and (*iii*) writing a small instruction workbook to guide the learners during their work (Kragten et al., 2021; Spitz et al., 2021b). If the intended lesson already exists, because it has been created and used before, the preparation becomes a simple selection step, often requiring only a few modifications of the available resources. Developing a new lesson is a more serious endeavor. Both, the construction of the workbook and the qualitative model (that is, the KB-norm) require advanced pedagogical and subject matter expertise and take a certain amount of time to create. Existing materials in terms of workbook templates and model patterns can be used to speedup this process. Templates and patterns also help to ensure quality.

Advanced explanation

Learners sometimes have subtle misunderstandings which are hard to overcome using logic-based explanations. For example, formal reasoning, as required by the learner when working with the qualitative representation, may get intertwined with confounding everyday concepts. This is where teachers make a significant difference. Consider the following. In a lesson on causes of poverty, a causal dependency represents the notion that an "increasing income will decrease poverty" (Figure 5). In the formal language this is represented using a negative causal dependency: the affected quantity (poverty) changes in the opposite direction of the causing quantity (income). However, in everyday conversation people typically say that "more income is good for poverty", implying that more income will improve the situation. We have observed that most learners, possibly after some "going back and forth", will grasp the correct interpretation. Acquiring this insight is actually a great learning experience and learners become better system thinkers. Yet, a small minority needs more advanced support that goes beyond the formal one. They often require a kind of dialogue that helps them to recognize and reflect on the misconception and guidance to revise their knowledge (Vosniadou et al., 2001). Compared to the current state of the technology a teacher is better at this task for two reasons. First, the required dialogue is advanced and often infused with specific knowledge concerning the learner involved. Second, the number of possible misunderstandings is potentially high and their kind is difficult to predict in advance. A teacher is typically more flexible and more able to address unexpected misconceptions as they occur.

Class management and learner engagement

There is set of tasks that are concerned with class management and keeping learners engaged. Often these tasks are not specific for the subject matter at hand, yet important for the learning experience to commence and ultimately be successful. It involves tasks such as welcoming learners, inquire about their wellbeing, ensuring a positive classroom climate, inspiring them to organize their materials and start working, and probably most important, keep learners engaged throughout the learning activity. Although intelligent software can vary a lot in terms of how motivating it is for learners, the overall tasks of class management and learner engagement are beyond the scope of current technology. Real time learning analytics can support the teacher identifying students who need attention. However, making sense of the learning analytics still needs to be done by the teacher because they are best informed about their students'

needs and the current situation in the classroom. Hence, these tasks remain with the teacher.

Conclusions and discussion

Although the potential of Artificial Intelligence software is widely proclaimed, its use in education is limited. To deploy this proclaimed potential, we investigate the use of a hybrid approach in which humans and intelligence software join forces in teaching. As a case study, we focus on learners acquiring systems thinking skills in secondary education and the new pedagogical approach that we are developing for this purpose. Different from typical problem solving tasks, in which case learners produce a particular answer (e.g., a number resulting from a calculation), learners use a knowledge representation (language), and an accompanying solver, and learn by creating a small knowledge base. The latter is presented to the learners as an interactive diagram.

After briefly reviewing and clarifying our understanding of what it takes to refer to a system as being an Artificial Intelligence system, we discuss the tasks best performed by such intelligent software and which tasks are best, or necessarily, given to the teacher. The presented approach is part of ongoing research, and both used and evaluated in real educational settings. There is a clear added value to this hybrid-approach because both "agents" can now excel in the tasks they are best at, which results in improved learning (Kragten et al.[2]).

Artificial Intelligence systems typically have an extensive storage of knowledge or information which they deploy when performing the task they were developed for. We take a different approach and work with small knowledge bases, often dedicated to a particular topic aligned with the subject matter that the teacher wants the learners to work on. Taking this approach is essential. Partly, because capturing *all* the required knowledge beforehand is simply not feasible. Moreover, having a dedicated knowledge base per lesson is very helpful in fine-tuning the rest of the interaction with the learner. Notice, that the approach is still generic and that all the interaction between the software and the learners is fully automated.

Being *explicit* is an important feature of the knowledge representation (language) central to the approach presented here. Firstly, because the concepts relevant to systems thinking are all explicitly represented as unique identifiable and tangible ingredients. This makes that learners work directly with the notions relevant to systems thinking, when they create their diagram and present it to the solver. The explicitness also facilitates the automated "agents" to directly read-off relevant information and deploy this in the interaction with the learner. As such, the problem of "explainable Artificial Intelligence" does not apply here, on the contrary.

Future work

Part of the steering during lessons in the classroom currently happens *via* the workbook. The workbook provides the learner textual information on the topic at hand, and has instructions about the steps to take. Part of the reason for having this workbook was the hypothesis that teachers prefer text-based instruments as being part of the overall setup. However, in the meantime experience in the classroom has shown that these documents create a certain amount of overhead. Teachers have requested if this can be handled in a different way. Hence, we are currently investigated whether the details provided in the workbook can also be automated, for instance based on the specifics of the knowledge base that we construct together with the teachers.

In the ongoing project, we work with a number of schools and their learners (K8-12). Each learner typically works with the approach presented multiple times per school year and over a number of consecutive years. As such, it is tempting to investigate the notion of a learner-model as a key component in the current set up. Having a learner-model would help to further tune the interaction to the specific needs of each individual student. However, the beauty of the current approach, thus without a learner-model, is that each student gets a fresh unbiased interaction each time. It is an open question whether the added value of a learner model would outweigh this benefit.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

2 Kragten, M., Spitz, L., and Bredeweg, B. *Learning Systems Thinking and Content Knowledge by Constructing Qualitative Representations in Lower Secondary Education*. (under review).

Publisher's note

References

Aicardi, C., Bitsch, L., Datta Burton, S., Evers, L., Farisco, M., Mahfoud, T., et al. (2020). *Trust and Transparency in Artificial Intelligence. Technical report D12.5.4, Human Brain Project SGA2*. Available online at: https://www.humanbrainproject.eu/ (accessed February 22, 2022).

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., et al. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 18–28. doi: 10.1109/MC.2020.2996587

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *Int. J. Artif. Intellig. Educ.* 26, 600–614. doi: 10.1007/s40593-016-0105-0

Ben-Zvi-Assaraf, O. B. Z., and Orion, N. (2010). System thinking skills at the elementary school level. *J. Res. Sci. Teach.* 47, 540–563. doi: 10.1002/tea.20383

BNVKI (2021). *Interview with Jaap van den Herik, Founding Father of the BNVKI*. Available online at: http://ii.tudelft.nl/bnvki/?p=1790 (accessed February 22, 2022).

Bouwer, A., and Bredeweg, B. (2010). Graphical means for inspecting qualitative models of system behaviour. *Instruct. Sci.* 38, 173–208. doi: 10.1007/s11251-008-9083-4

Bratko, I. (2012). *Prolog Programming for Artificial Intelligence, 4th Edn*. Wokingham: Addison-Wesley.

Bredeweg, B., Liem, J., Beek, W., Salles, P., and Linnebank, F. (2010). "Learning spaces as representational scaffolds for learning conceptual knowledge of system behavior," in *Technology Enhanced Learning, LNCS 6383*, eds M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, and V. Dimitrova (Heidelberg: Springer), 46–61.

Bredeweg, B., Linnebank, F., Bouwer, A., and Liem, J. (2009). GARP3 - Workbench for qualitative modelling and simulation. *Ecol. Inform.* 4, 263–281. doi: 10.1016/j.ecoinf.2009.09.009

Bredeweg, B., and Salles, P. (2009). Qualitative models of ecological systems (Editorial introduction). *Ecol. Inform.* 4, 261–262. doi: 10.1016/j.ecoinf.2009.10.001

Chou, C. Y., Huang, B. H., and Lin, C. J. (2011). Complementary machine intelligence and human intelligence in virtual teaching assistant for tutoring program tracing. *Comput. Educ.* 57, 2303–2312. doi: 10.1016/j.compedu.2011.06.005

Curriculum.nu (2021). Available online at: https://www.curriculum.nu (accessed February 22, 2022).

Davis, R., Shrobe, H., and Szolovits, P. (1993). What is a knowledge representation? *AI Magazine* 14, 17–33.

Forbus, K. D. (2018). *Qualitative Representations. How People Reason and Learn About the Continuous World*. Cambridge, MA: The MIT Press.

Gentner, D., and Stevens, A. (1983). *Mental Models*. New York, NY: Lawrence Erlbaum Associates.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Holstein, K., Aleven, V., and Rummel, N. (2020). A conceptual framework for human–AI hybrid adaptivity in education. *Artificial Intelligence in Education, LNAI* 12163, 240–254. doi: 10.1007/978-3-030-52237-7_20

Holstein, K., McLaren, B. M., and Aleven, V. (2019). Designing for complementarity: teacher and student needs for orchestration support in AI-enhanced classrooms. *Artif. Intellig. Educ. LNAI* 11625, 157–171. doi: 10.1007/978-3-030-23204-7_14

Jacobson, M. J., and Wilensky, U. (2006). Complex systems in education: scientific and educational importance and implications for the learning sciences. *J. Learn. Sci.* 15, 11–34. doi: 10.1207/s15327809jls1501_4

Kragten, M., Spitz, L., and Bredeweg, B. (2021). "Learning domain knowledge and systems thinking using qualitative representations in secondary education (grade 9-10)," in *Proceedings of the 34th International Workshop on Qualitative Reasoning* (Montreal, QC).

Kuhn, M., and Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. New York, NY: Chapman and Hall/CRC.

Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. New York, NY: Viking Books

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Liem, J. (2013). *Supporting conceptual modelling of dynamic systems: A knowledge engineering perspective on qualitative reasoning* (PhD thesis). University of Amsterdam, Netherlands.

Lindsay, P. H., and Norman, D. A. (1977). *Human Information Processing - An Introduction to Psychology, 2nd Edn*. New York, NY: Academic Press.

Marcus, G., and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We can Trust*. New York, NY: Pantheon Press.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Available online at: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed February 22, 2022).

McCullogh, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. doi: 10.1007/BF02478259

Moschoyiannis, S., Peñaloza, R., Vanthienen, J., Soylu, A., and Roman, D. (eds.) (2021). "Rules and reasoning," in *5th International Joint Conference RuleML+RR, LNCS 12851* (Heidelberg: Springer).

NGSS (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Nirenburg, S. (2017). Cognitive systems: towards human-level functionality. *AI Magazine* 38, 5–12. doi: 10.1609/aimag.v38i4.2760

Paiva, R., and Bittencourt, I. I. (2020). Helping teachers help their students: a human-AI hybrid approach. *Artif. Intellig. Educ. LNAI* 12163, 448–459. doi: 10.1007/978-3-030-52237-7_36

Prain, V., and Tytler, R. (2012). Learning through constructing representations in science: a framework of representational construction affordances. *Int. J. Sci. Educ.* 34, 2751–2773. doi: 10.1080/09500693.2011.626462

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi: 10.1037/h0042519

Rosenshine, B. (2012). Principles of instruction: Research-based strategies that all teachers should know. *Am. Educ.* 36, 12–39.

Russell, S., and Peter Norvig, P. (2020). *Artificial Intelligence: A Modern Approach, 4th Edn*. Hoboken, NJ: Pearson.

Schreiber, G., Wielinga, B., and Breuker, J. (1993). *KADS: A Principled Approach to Knowledge-Based System Development*. London: Academic Press.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270

Skinner, B. F. (1974). *About Behaviorism*. New York, NY: Vintage Books.

Spitz, L., Kragten, M., and Bredeweg, B. (2021a). "Exploring the working and effectiveness of norm-model feedback in conceptual modelling: a preliminary report," in *International Conference on Artificial*

Intelligence in Education, LNCS 12749, eds I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova (Heidelberg: Springer), 325–330.

Spitz, L., Kragten, M., and Bredeweg, B. (2021b). "Learning domain knowledge and systems thinking using qualitative representations in secondary education (grade 8-9)," in *Proceedings of the 34th International Workshop on Qualitative Reasoning* (Montreal, QC).

Sweeney, L. B., and Sterman, J. D. (2007). Thinking about systems: student and teacher conceptions of natural and social systems. *Syst. Dynam. Rev.* 23, 285–311. doi: 10.1002/sdr.366

van Harmelen, F., Lifschitz, V., and Porter, B. (eds.) (2008). *Handbook of Knowledge Representation*. Amsterdam: Elsevier.

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., and Papademetriou,. E. (2001). Designing learning environments to promote conceptual change in science. *Learn. Instruct.* 11, 381–419 doi: 10.1016/S0959-4752(00)00038-4

Weld, D. S., and de Kleer, J. (1990). *Readings in Qualitative Reasoning About Physical Systems*. San Mateo, CA: Morgan Kaufmann.

Wenger, E. (1987). *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge*. Los Altos, CA: Morgan Kaufmann.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 25 August 2022
DOI 10.3389/frai.2022.910801

OPEN ACCESS

EDITED BY
Julita Vassileva,
University of Saskatchewan, Canada

REVIEWED BY
Riccardo De Benedictis,
National Research Council (CNR), Italy
Emmanuel G. Blanchard,
IDU Interactive Inc., Canada

*CORRESPONDENCE
Alice Plebe
alice.plebe@unitn.it

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 01 April 2022
ACCEPTED 10 August 2022
PUBLISHED 25 August 2022

CITATION
Plebe A, Rosati Papini GP, Cherubini A and Da Lio M (2022) Distributed cognition for collaboration between human drivers and self-driving cars. *Front. Artif. Intell.* 5:910801. doi: 10.3389/frai.2022.910801

Distributed cognition for collaboration between human drivers and self-driving cars

Alice Plebe*, Gastone Pietro Rosati Papini, Antonello Cherubini and Mauro Da Lio

Department of Industrial Engineering, University of Trento, Trento, Italy

This paper focuses on the collaboration between human drivers and intelligent vehicles. We propose a collaboration mechanism grounded on the concept of distributed cognition. With distributed cognition, intelligence does not lie just in the single entity but also in the interaction with the other cognitive components in a system. We apply this idea to vehicle intelligence, proposing a system distributed into two cognitive entities—the human and the autonomous agent—that together contribute to drive the vehicle. This account of vehicle intelligence differs from the mainstream research effort on highly autonomous cars. The proposed mechanism follows one of the paradigm derived from distributed cognition, the *rider-horse* metaphor: just like the rider communicates their intention to the horse through the reins, the human influences the agent using the pedals and the steering wheel. We use a driving simulator to demonstrate the collaboration in action, showing how the human can communicate and interact with the agent in various ways with safe outcomes.

KEYWORDS

autonomous driving, distributed cognition, human-vehicle collaboration, human-robot interaction, emergent behavior, artificial intelligence

1. Introduction

Recent developments in autonomous driving are leading to a transitional period, where human drivers and intelligent vehicles coexist. Nowadays, more and more commercial vehicles feature intermediate levels of automation. The presence of partially autonomous vehicles on the streets is starting to affect the traditional driver-vehicle interaction patterns. In fact, the addition of automation leads to a significant behavioral change in the way humans drive; interacting with partially automated systems disrupts the classic traffic dynamics, and it can cause unsafe interactions difficult to predict (Flemisch et al., 2017). Hence, the research community must place at the top of its agenda the issue of cognitive interaction between the driver and the automated system.

To date, research on vehicle intelligence has mainly addressed fully autonomous cars. They are far from the idea of human-vehicle collaboration, because the greater the automation, the less the human is involved in the driving task. In fact, the ideal self-driving vehicle would dispense with the human and any form of collaboration

with them. The account of vehicle intelligence completely separated from the human driver has developed considerably, also because of the ongoing evolution of deep learning. However, the research is still far from achieving totally driverless vehicles, and it often overlooks the importance of mutual dependence between the human driver and the vehicle.

We argue that new forms of collaboration between humans and artificial agents can arise from the theoretical framework of distributed cognition, i.e., the idea to achieve a task through the emergent interaction of more intelligent entities. In the effort to achieve artificial driving agents with increasingly cognitive abilities, we see a promising direction in the idea of a distributed cognitive system: two cognitive entities—the human and the agent—collaborate to achieve the task of driving the vehicle. As we will show, this framework promotes new interesting ways to approach human-agent collaboration, leading to the formulation of a number of "metaphores" suggesting ideal styles of interaction.

We present a collaboration paradigm showing the advantages of having a system with more than a single cognitive entity. The system follows the *rider-horse* metaphor to implement distributed cognition. As the horse can "read" human's intentions and, reciprocally, the rider can understand animal's intentions, we argue that autonomous vehicles might benefit from a similar ability: the user experience would improve if the driver could give hints to the car and feel as if the car could "understand" their intentions. While the rider-horse system communicates with the reins, the human communicates with the agent using the pedals and the steering wheel. We show the collaboration system in action on a driving simulator. The results illustrate how the human can influence the agent's decision-making to obtain, for example, a lane change or an overtake whenever possible and safe; on the other hand, the agent can dismiss the human's suggestion if they are dangerous or not significant.

The following Section briefly introduces the different accounts of cognition proposed through the years, focusing especially on the distributed nature of cognition. Section 3 presents the main research direction pursued in autonomous driving, which sets aside the idea of collaboration with the human and focuses on vehicle intelligence as single cognition. Section 4 dives into the distributed account of vehicle intelligence and analyzes various collaboration paradigms between humans and autonomous agents. Section 5 presents the interaction mechanism we propose between a human and an autonomous agent previously developed. The section describes how the agent works (in brief) and how the interaction mechanism is realized. Section 6 demonstrates the system in action using a driving simulator. Lastly, Section 7 draws the conclusion and discusses future work.

2. Accounts of cognition

It already exists a form of intelligence capable of driving vehicles—humans. Thus, it is reasonable to design other forms of "vehicle intelligence" by taking inspiration from human intelligence and cognition. Human cognition is the focus of a vast area of research with a long-stand history. It is useful here to briefly sketch the different accounts of cognition proposed through the years, with special attention to the distributed nature of cognition.

One of the core ideas of cognitive science, at the time of its birth in 1956, is that minds and computers are exemplars of the same class, the *physical symbol system* (Gardner, 1985). A fundamental corollary of this theory is that what is possible for a human mind—for example, driving—is possible for a computer as well. This idea works in principle, but there is still no clear understanding on what kind of computations the human brain runs when, again for example, the person is driving a car.

Cognition has been characterized with a distributed structure since the early period of physical symbol. Newell and Simon (1972) proposed an abstract structure divided into perceptual modules, a central information processing system, and motor activation modules. Fodor (1983) postulated that the mind is a collection of autonomous modules, with independent information, communicating by input/output. Moreover, Minsky (1986) proposed that the mind is like a society, in which each inhabitant has their own job and cooperates with the rest toward common goals. Rumelhart and McClelland (1986) highlighted how cognition is distributed over a large number of interconnected units. The "distributed" account of cognition mentioned so far considers only a single cognitive entity, composed of several sub-parts. However, there is cognition beyond the individual intelligence.

The idea of *distributed cognition* became popular with the work of Hutchins (1995b,a). This current of though stresses how the highest cognitive functions imply a strong social relation and cannot be studied in isolation (Cole and Engeström, 1993). Hutchins founded the concept of distributed cognition on his extended cognitive ethnography of ship navigation (Hutchins, 1995a): a ship requires a complex system made by both a team of people and an array of technologies, all working together. The team is organized, with precise roles for each crew member, and the cognitive work is offloaded thanks to aids such as instruments and charts. Hutchins further extended his study of distributed cognition from ship navigation to aviation—a domain closer to the focus of this paper than sea navigation (Hutchins, 1995b). The case analyzed by Hutchins is the management of the airplane's speed during landing. Speed is the most crucial factor for a safe landing. The process involves coordination within the crew as well as interaction with the instrumentation. Hutchins' account of cognition has been accepted as the best way to describe the dynamics

and complexity of various human organizations, including classrooms, office work, company organization, and air traveling (Dror and Harnad, 2008).

Despite the innovation of Hutchins' work, cognitive science of that time was dominated by another school of though, called *4E Cognition* (Newen et al., 2018). The "4E" approach characterizes cognition with four features: *embodied*, *embedded*, *enacted*, and *extended*. Embodied cognition analyzes not just the role of the mind, but also the role the body has in cognition. Embedded cognition focuses on the integration of the cognitive agent into an environment. Enacted cognition assumes that knowledge is closely related to the notion of action (for example, perception is not just something propaedeutic to an action, it is a sort of action itself). We will not go in detail of these accounts of cognition as they are not the focus of the discussion; it is the last "E," in fact, the most relevant in our context.

Extended cognition (Clark and Chalmers, 1998; Clark, 2008) presents an even more radical account of cognition than the one proposed by Hutchins. Extended cognition accepts as active components of cognition all kinds of things that can help humans think. Clark's famous example is the notebook used by a person suffering from memory loss. The person uses the notebook to take note of everything they need to know. For the person's cognition, the notebook plays a role as crucial and constitutive as their biological memory. Unlike Hutchins's, Clark's proposal spurred a huge debate within cognitive science (Menary, 2010), and it has become a key theoretical framework for topics such as the *Internet enhancement of cognition* (Smart, 2017). However, in the context of vehicle intelligence, Clark's notion of extended cognition is not so apt. According to him, the cognitive system gives equal partnership to the human mind and the external component. This aspect is questionable when the external part is trivially poor from the cognitive point of view—like the notebook example—or when the external part is a knowledge-packed resource like Wikipedia or Google.

The cooperative relation between humans and their extended cognitive counterpart is well-represented in the framework proposed by Poirier and Chicoisne (2008). They present a two-dimensional conceptual space (see Figure 1) to classify the cooperation between two entities in a cognitive system. One axis represents the degree of cognition of the entities, ranging from *cognitive* to *non-cognitive*. For example, a pencil is totally non-cognitive, while humans have maximum degree of cognition. The other axis represents the outcome of the cooperation, which ranges from *aggregate* to *emergent*. Aggregate means there are no cooperative or inhibitory interactions among the parts of the system, and the task can be achieved even when parts of the system are removed. Emergent represents the opposite, when the parts of the system collaborate actively to achieve a shared task. For example, two researchers write a scientific a paper, and they agree to split the work in half; each person writes only a specific section of the paper without reading the rest. Only at the end, they merge the sections together. In this way, there is no cognitive advantage from the cooperation, because it is a simple aggregation of individual cognitive loads. Although the entities have high degree of cognition, the collaboration is aggregate. Any approach exploiting the concept of distributed cognition should fall in quadrant 2 of this conceptual space.

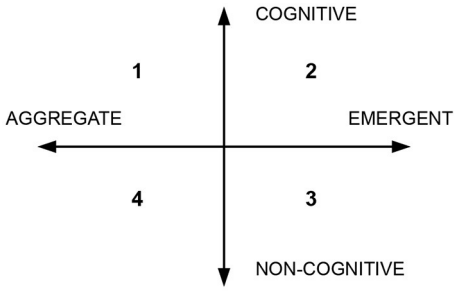

FIGURE 1
Conceptual space proposed by Poirier and Chicoisne (2008) to classify the relation between two entities in a cognitive system.

In the following sections, we will analyze current approaches to vehicle intelligence and how they relate to distributed cognition.

3. Vehicle intelligence with single cognition

The main research direction in autonomous driving focuses on developing high levels of driving automation—the higher the level, the less the human is involved in the driving task. The Society of Automotive Engineers (SAE) defines six levels of driving automation (SAE, 2021), summarized in Figure 2. Level 0 stands for no automation at all, i.e., traditional automobiles. Level 1 introduces basic forms of driver assistance, such as emergency braking. Level 2, also called *partial automation*, is the form of automation currently available on recent vehicles, and it includes systems like adaptive cruise control and lane following. However, the human is still responsible of driving the car and must constantly supervise the system. Level 3, also called *conditional automation*, introduces a drastic shift from the previous level. Here, the system is responsible for driving the car and supervising the scene, while the human is allowed to engage in other activities. These systems operate in limited operational design domains, usually highways. Still, emergency situations might occur where the system is not able to proceed safely: in these cases, the system disengages from the driving task and requires the human to resume control of the vehicle with short time. Level 4 does not need human supervision. The system can work even without a person inside the vehicle. However, it still operates in limited domains.

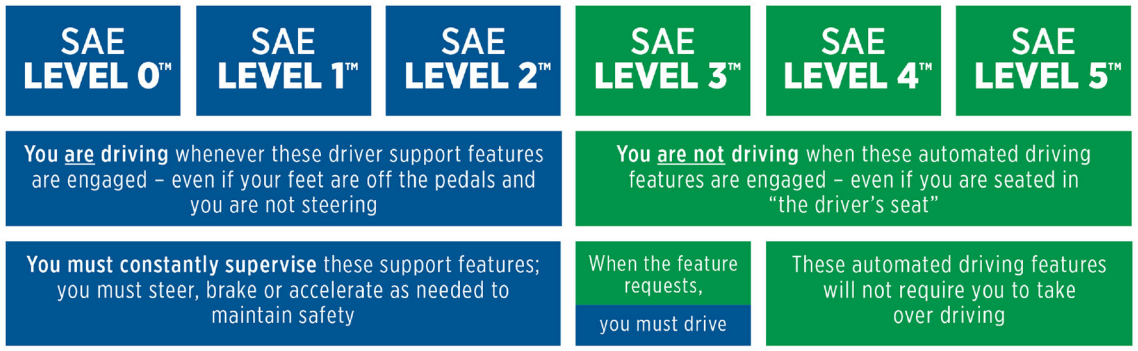

FIGURE 2
SAE levels of automation SAE (2021).

The operational domains usually consist of highway scenarios, which are easier to manage with respect to urban scenarios. Driving in urban areas presents a bigger challenge because of multiple traffic directions, intersections, parked vehicles, traffic lights, sidewalks, and numerous classes of *vulnerable road users*. Lastly, Level 5 represents full automation—ideally, a car without steering wheel and pedals. Here, the human driver is completely replaced by the system, which is able to operate in any conditions without limitations.

Most of the research effort is now put into developing Levels 4 and 5. This research direction overlooks the idea of collaboration between human and driving agent. The vehicle intelligence, in this account, aims to gradually assume the role of the driver and make the human simply a passenger not involved in the driving task. In fact, the higher the level of autonomy, the more the human driver is replaced by the artificial agent. This approach to autonomous vehicles is far from distributed cognition—there are two cognitive entities in the system, but there is no collaboration between them. Either the agent or the human is in charge of controlling the vehicle, and when necessary the control passes to one another. Disengagements, i.e., where the human must resume control of the vehicle because the agent stops working safely, are one of the most critical aspects of autonomous driving systems. The problem of disengagements affects Levels 3 and 4 the most, precisely because the collaboration between the cognitive entities in the system is missing. In fact, the more automation is added to the system (and the more reliable and robust the autonomy is), the human is less likely to predict the automation failure dues to the lack of cognitive engagement (Endsley, 2017). For this reason, Levels 3 and 4 are paradoxically less reliable than Level 2, where the human should be constantly supervising the system (however human beings tend to misuse Level 2 not supervising as they should).

Level 5 of autonomy can be the solution to the conundrum of disengagements, since the system would never require the person to take over driving. Completely replacing human drivers with artificial drivers is indeed desirable, but still a challenging task. Production-level deployment of full self-driving vehicles remains a distant future (Jain et al., 2021). On the one hand, state-of-the-art driving agents surpass humans in computation, responsiveness, and multitasking. On the other hand, humans exceed automation in the capacity of detection, context understanding, induction, and improvisation (Xing et al., 2021). For this reason, researchers are looking at new directions to develop Level 5 systems focusing on cognitive-inspired approaches. To achieve an AI capable of handling any possible (or unseen) traffic scenario, it appears more and more necessary to develop high-level cognitive abilities similar to humans (Wang et al., 2021). Implementing human-like cognitive behaviors is far from easy. As discussed in Section 2, there are countless theories trying to progress the understanding of the mind and the brain. The current understanding of how the brain executes complex behaviors such as driving is vague, often controversial, and short of detail.

Given the challenges linked to Level 5 systems, a parallel research direction looks at the concept of distributed cognition applied to vehicle intelligence. The idea is to design systems where the collaboration between human and agent is at the core of the driving mechanism. This approach takes the best of both worlds, leveraging the potential of human intelligence and the computational power of machine intelligence.

4. Vehicle intelligence with distributed cognition

As mentioned in the Section 1, vehicle automation will cause significant behavioral changes in human driving. The behavioral change depends on the way vehicle intelligence is designed. In the "single cognition" account reviewed in Section 3, humans are gradually removed from the driving task. However, the

transition from Levels 2–3 to Levels 4–5 is proceeding slowly, forcing human drivers to interact with partially automated systems—often without being aware that other vehicles are controlled by artificial agents. These interactions disrupt the classic traffic dynamics and can produce unsafe scenarios (e.g., disengagements) that are difficult to predict (Flemisch et al., 2017).

The "distributed cognition" account of vehicle intelligence approaches the problem of driver-vehicle interaction patterns differently. Cooperative vehicle intelligence is grounded on the idea that, in a system, knowledge does not lie solely within the individual but rather within all entities involved in the system (Banks and Stanton, 2017). This follows Hutchins' account of distributed cognition, described in Section 2. Applying this idea to intelligent vehicles means that humans and driving agents must collaborate actively. The driving task is achieved only through the interaction of the two entities, because each contributes with a different (if not complementary) set of cognitive skills.

It is not straightforward to determine how the skills of drivers and automated vehicles can be combined for optimal cooperation. Researchers have proposed metaphors to extract design concepts for ideal human-agent interaction. Marcano et al. (2020) pinpoint four metaphors used as "blueprint" for distributed driving systems. The first is the *rider-horse* metaphor, also called *H-Metaphor* (Flemisch et al., 2003). It compares the human-agent interaction to a human riding a horse. When riding, the human controls the horse through the reins. This haptic interface allows the horse and the rider to "understand" each other's intentions. In addition, the rider can take the horse under tight reins to exert more direct control or can use loose reins to provide the horse with a higher degree of autonomy. The second metaphor is the *aviator instructor-student* (Holzmann et al., 2006). It describes the interaction occurring in a flying training session between a student and an experienced aviation pilot. The expert aviator assists the beginner either actively (by exerting forces on the control system) to help with the execution of maneuvers, or passively (by holding the steering control with different forces) to approve or disapprove the student's action. The next metaphor is the *joint-carrying* of an object (Flemisch et al., 2016). It emphasizes the collaboration between two agents that share the same task and interact physically on the same object. The interesting aspect is that the agents have different perception capabilities—in the specific example, one is walking forward and the other backwards. Yet, the information perceived by an agent complements each other, and both are needed to complete the task. Lastly, the *parent-child* metaphor illustrates a parent teaching a child to ride a bike (Flemisch et al., 2012). In this metaphor, the child has control of the bike, and the parent does not interfere while the child is performing well. If the child starts wobbling, the parent intervenes in proportion to the risk—the intervention should be gentle in any case, to avoid rejection of the assistance.

All metaphors are relevant to the case proposed here, but with various degrees. The least relevant metaphor is the *parent-child*, while it is certainly true that the autonomous system should avoid to overwhelm drivers while they are performing well, and gently intervene if the driver leads the vehicle to an unsafe condition. The *joint-carrying* metaphor describes well one specific aspect: the different and complementary perceptions of the scene by the driver and the system. However, it goes no further in indicating how these differences should be reconciled. The *aviator instructor-student* metaphor brings us back into the domain of aviation, which has certain affinities with autonomous driving, as commented in Section 2. Aeronautics has a long history of automated procedures and human-computer interactions. However, there are obvious differences with respect to autonomous driving. For example, in the context of airplanes, distributed cognition implies a distribution of roles within the crew, while this is irrelevant in an autonomous car. Moreover, a vehicle continuously interacts with the environment and the other road users at close range. On the other hand, there are lessons that can be taken from the field of aviation. As the role of the driver becomes gradually closer to that of an airplane pilot, a new class of errors can lead to incidents. In aviation, a *classification error* occurs when the pilot assumes that the system is working in a way that is different from the actual state of the system. This form of error seems likely to occur within driving automation as well—this is discussed in more detail in Banks and Stanton (2017, p. 15–16). It is, however, the *rider-horse* metaphor that captures in the best and most complete way the current proposal, as we will explain in Section 5.

Reviews on driver-vehicle collaboration can be found in Xing et al. (2021), Marcano et al. (2020), Bengler et al. (2014), and Michalke and Kastner (2011). Works focus on key factors like human trust and situation awareness, which influence the design of the system. Moreover, the form of interaction defines the control mechanism—we can distinguish between shared control and take-over control. The type of control mechanism determines also how to implement the steering/pedal system, either with a coupled or uncoupled control framework. However, not all attempts at driver-vehicle collaboration can be considered forms of distributed cognition. Recalling the diagram of Figure 1 proposed by Poirier and Chicoisne (2008), there are approaches that fall outside quadrant 2, which is the only quadrant identifying distributed cognition. Consider, for example, low-level ADAS systems such as emergency brake or lane departure warning: they have very low degree of cognition. Hence, it is not possible to talk about distributed cognition—they belong to quadrant 4. On the other hand, more advance (cognitive) systems like overtaking assistance tends to generate aggregate outcomes, as opposite of emergent outcomes according to the classification of Poirier and Chicoisne (2008). In these systems, there is no overt cooperative interactions between the human and the assistant: the assistant is either on or off, and there is no mean of communication between the parts.

Moreover, when the assistant is off, the human can still achieve the driving task. Therefore, the systems are not emergent and fall in quadrant 1.

In the next section, we will describe our implementation of distributed cognition in an autonomous driving system.

5. Methodology

We present a collaboration paradigm between a self-driving agent and a human driver based on the H-metaphor. In the proposed system, just like the rider influences the horse's behavior using the reins, the human driver steers the decision-making of the autonomous agent using the pedals and the steering wheel. The system uses an uncoupled control framework and is an example of distributed cognition applied to vehicle intelligence. The two entities in the system are both intelligent, both interpreting the world, and working jointly to achieve the same task. The self-driving agent has high degree of cognition and collaborates with the human in an emergent way. Therefore, the proposed systems falls in quadrant 2 of the classification space of Poirier and Chicoisne, in Figure 1.

The autonomous agent considered here has been developed within the European H2020 project Dreams4Cars[1]. The agent has the cognitive capabilities necessary to drive a vehicle autonomously, in controlled situations, and it can be regarded as Level 4 in the SAE definition. Note that this work focuses on the interaction paradigm between the agent and the human, rather than how the autonomous agent works. Here, we include only a brief explanation of the agent to better understand the collaboration mechanism; a detailed description of the agent architecture is in Da Lio et al. (2020).

The sensorimotor system of the agent is designed to be compatible with the human system. Specifically, the agent must be capable of seeing the action possibilities latent in the environment—dubbed *affordances* by Gibson (1986)—and it has to generate the corresponding action plans in a way similar to a human driver. An example of affordance, taken from Gibson's original work, is the vision of a stair; it elicits the action of stepping, up or down, relative to the size of the person's legs. Another example—very close to the problem under consideration here—is the following: *"The progress of locomotion is guided by the perception of barriers and obstacles, that is, by the act of steering into the openings and away from the surfaces that afford injury"* (Gibson, 1986, p. 132). For a self-driving agent, the affordances are the physically traversable space constrained by traffic rules and space-time restrictions from moving obstacles. The rider-horse collaboration has the same scheme, since the horse sees the same affordable paths of the rider, and the rider can infer the horse's intentions.

1 https://www.dreams4cars.eu

The agent works in two phases: action priming and action selection. During action priming, the agent detects the set of affordances D in the navigable space and maps them onto estimates of their *salience*. The salience measures how good the corresponding action is. The actions that the agent can produce are the set of trajectories U (i.e., time-space locations of the vehicle) that originate from the current configuration. Since a vehicle has two controllable degrees of freedom, the whole space of possible actions is spanned by the specification of the longitudinal and lateral controls. In our implementation, the longitudinal control is the jerk j, and the lateral control is the steering rate r (i.e., the time derivative of the steering angle). For an instantaneous action $u = \langle j, r \rangle$, $\nu(u, d)$ represents how good or desirable the action is in relation to the affordance d. $\nu(u, d)$ evaluates two factors: the probability of remaining in the specified spatial domain of the affordance d for a sufficient time; the travel time subject to speed limits and comfort criteria. The salience to express how good the choice of the current control $\langle j, r \rangle$ for the affordance d is the following:

$$s_d(j, r) = \sup_{u \in U} \{\nu(u, d)\}. \tag{1}$$

This means that the salience of the instantaneous choice $\langle j, r \rangle$ for the affordance d is the value $\nu(\tilde{u}, d)$ of the optimal action $\tilde{u}$ among all actions beginning with $\langle j, r \rangle$. The global salience function can be defined as follows, where weights w_d serve to prioritize sets of affordances:

$$s(j, r) = \max_{d \in D} \{w_d \, s_d(j, r)\}. \tag{2}$$

During the second phase of action selection, the agent chooses the motor control $\langle j, r \rangle$ corresponding to the maximum salience and executes it.

The way the autonomous agent works is broadly inspired by how human cognition (presumably) realizes the driving task. Hence, it is reasonable to expect that a similar process occurs in the mind of the human inside the vehicle the agent is controlling. The human recognizes their own set of affordances and computes a salience value $s^*(j, r)$ for each action they have in mind. We assume that the sensorial system of the vehicle is reliable—as it is indeed in most situations—and that the system has learned an efficient control policy in response to affordances. Hence, is it reasonable to expect in most cases that $s^*(j, r) \approx s(j, r)$. For a more detailed explanation, see Da Lio et al. (2017). However, there can be situations where the person desires a different action or have a specific goal in mind the agent is not aware of. With distributed cognition, the human can obtain the desired behavior by collaborating with the agent, which is able to interpret the human's intention.

The human interacts with the agent by biasing the action selection process, through the pedals and the steering wheel. The gas/brake pedals control the longitudinal bias, and the steering wheel controls the lateral bias. The biases influence

10.3389/frai.2022.910801

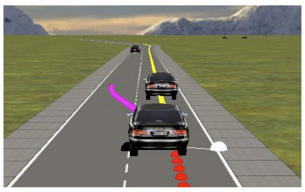

FIGURE 3
Screenshot of the OpenDS simulator showing the human biasing the agent to overtake. The scenario corresponds to Figure 4(i), showing the affordances a in yellow and b in purple.

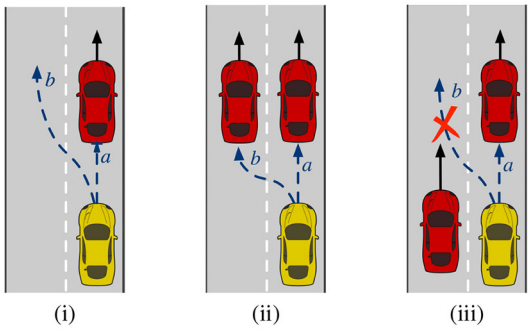

FIGURE 4
Three simulated scenarios to test the collaboration between the agent and the human (yellow car) when other vehicles are present (red cars). The dashed arrows show the affordable actions. **(i)** It is possible to follow the red car or to overtake. **(ii)** It is possible to follow the car on the right or the one on the left. **(iii)** The only affordable action is to follow the car ahead.

the computation of the global salience (2) by applying weights either to sets of affordances or to individual ones. In the case of longitudinal bias, the human can suggest the agent to drive faster/slower by pressing the gas/brake pedal—hence, applying a weight to the faster/slower affordances. The modified salience function is the following:

$$s'(j,r) = k(g-b)j\, s(j,r) \tag{3}$$

where g and b are the normalized gas and brake strokes, and k is a convenient gain. In the case of lateral bias, the human can prompt the agent to change lane to the left/right by steering the wheel. This action weights the individual affordances corresponding to lane change in the suggested direction.

The presented collaboration paradigm works safely because of distributed cognition. Since the system is composed of two cognitive capable entities, each of them can supervise the other and prevent wrong behaviors. For example, if the human suggests to perform a dangerous or unfeasible maneuver, the agent ignores the command. In fact, the agent dismisses any action that is not affordable or for which the salience is low or inhibited. This mechanism is one of the most critical part in the system; Section 7.1 further analyzes its limitations and how to resolve them. The next section provides simulations demonstrating these safe behaviors.

6. Demonstrations

We test the collaboration mechanism in the open-source driving simulator called OpenDS[2], depicted in Figure 3. This Section describes the outcome of five tests carried out in three simulated scenarios. In each test, we focus on how the human can promote various driving actions by collaborating with the autonomous agent.

The first test scenario shown in Figure 4(i) is a two-lane straight road where overtaking is possible and safe. In the same lane, there are the ego-vehicle and another car ahead of it, colored in yellow and red respectively. The autonomous agent identifies two possible affordances: to follow the red car (affordance a in the figure), or to overtake (affordance b). Since the leading vehicle is driving at almost the speed limit, the agent chooses the affordance a. When the human steers the wheel to the left, they exert a bias toward the affordances corresponding to the left lane (just b in this example). As a result, the salience of affordance b surpasses a, and the agent executes the action of overtaking the red car. A second test with the same scenario demonstrates the same outcome but with a different form of interaction. This time, the human promotes the overtake by pressing the gas pedal rather than steering the wheel. The positive bias affects the weights of faster affordances, that is b. Hence, just like before, the agent shifts from a to b and overtakes the leading car.

In the next scenario, Figure 4(ii), there are two cars ahead of the ego-vehicle, which occupy both lanes and travel at the same speed. In this case, affordances a e b have the same longitudinal control, i.e., b is not faster than a as before. The salience of a is grater than b because the latter discounts the cost of changing lane. Hence, the agent chooses a. If the human steers the wheel to the left, choosing the affordance b does not lead to any tangible speed improvement. However, the agent understands the human's desire and moves to the adjacent lane and starts following the left car. Using the same scenario, another test shows what happens if the human tries to bias the agent using the gas pedal rather than the steering wheel. In this case, the human bias has no effect because there are no affordable faster actions. The cars ahead do not allow the agent to increase the speed. Therefore, the human request cannot be satisfied, and agent keeps affordance a.

2 https://opends.dfki.de/

The last scenario, Figure 4(iii), shows a car ahead of the ego-vehicle and a car overtaking on the left lane. Here, affordance b no longer exists: the car on the left prevents the agent from changing lane. Since there are no affordable actions linked to the left lane, when the human steers the wheel of presses the gas pedal, there is no effect. The agent ignores the human's request and remains on the right lane following a. Further simulations are available in Da Lio et al. (2022).

7. Discussion

In this paper, we have argued that vehicle intelligence can benefit from the theoretical concept of distributed cognition. Distributed cognition can help designing new paradigms of collaboration between human drivers and autonomous agents. Cooperative vehicle intelligence is grounded on the idea that, in a system, knowledge does not lie solely within the individual but rather within all entities involved in the system. This research line moves away from the mainstream development of autonomous driving, which aims to completely remove humans from the driving task. However, fully autonomous vehicles are still far from being achieved, while distributed vehicle intelligence can solve at the present time the problems caused by the disruption of classical traffic dynamics.

We have proposed a collaboration paradigm founded upon the rider-horse metaphor, allowing the human to influence the decision making of the driving agent. Just like the rider communicates their intention to the horse through the reins, the human interacts with the agent using the pedals and the steering wheel. If the human presses the gas/brake pedal, they suggest the agent to drive faster/slower. If the human steers the wheel, they suggest the agent to change lane.

The collaboration system can support the user also in situations where the human is normally uncertain on how to behave. For example, the user hesitates to overtake because they are not sure about the feasibility of the maneuver. The user may want to drives closer to the opposite lane to see better ahead before deciding whether to overtake. With our collaboration mechanism, the user is not responsible to evaluate if it is safe or not to overtake. It is the agent that performs the evaluation and, in positive case, executes the overtake. Therefore, there is no need anymore for the user to drive for a moment to the center to have a clearer view of the road, because the agent is the one responsible to check if the overtake is feasible. Even if the user steers to the side to see ahead, the agent will not execute the overtake if it deems the maneuver risky (note that the perception system of the agent differs from the human, so the agent does not actually need to drive to the side to see better, like a human driver would do—although there are new attempts at human-inspired perception for autonomous vehicles; Plebe et al., 2021). This collaboration paradigm leads to a new way of driving, and human drivers will need some time to adjust to it. With this collaborative style of driving, humans, and agents become responsible for decisions at different levels: the agents take care of the execution of safe maneuvers, and the humans decide on the overall driving style, e.g., faster/slower, conservative/aggressive, and such.

7.1. Limitations and future work

The distributed cognition approach works best when the entities in the system lie close in the "cognitive" axis of the classification space of Figure 1. In the context of vehicle intelligence, this means that the human and the driving agent should be capable of understanding each other's intentions. In other words, they should share the same affordances. Unfortunately, this is not always the case. Fully autonomous and reliable driving agents do not exist, yet. Hence, unpredictable situations are still possible, where the human detects unconventional affordances that the autonomous agent is not aware of. For example, if the road is blocked by a tree, an autonomous agent would stop forever; however, a human driver could be aware that the ground on the side of the road is "driveable" and that it is possible to bypass the blockage by driving on the gravel. This is an affordable action that the agent would most likely miss.

Future work is to extend the collaboration paradigm with a "tight reins" mode—to use an expression in accordance with the H-metaphor. With this mode, the human can apply a tight control to make the agent accepts affordances not known before and generate new behaviors. If the user insists on an action that the agent is refusing to perform because not affordable in its view, after a certain "persistence threshold", the agent accepts the new affordance and executes the maneuver. A similar concept can be found in Vanholme et al. (2011) for driving on highways. This solution would mitigate the issue of artificial systems dangerously overriding human decisions—an issue common also to other research domain such as aeronautics.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

AP wrote the manuscript with input from all authors. GR and AC carried out the simulations. MD was in charge of overall

direction and planning. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Banks, V. A. and Stanton, N. A. (2017). *Automobile Automation: Distributed Cognition on the Road (1st ed.)*. Boca Raton, FL: CRC Press. doi: 10.1201/9781315295657

Bengler, K., Dietmayer, B., Maurer, M., Stiller, C., and Winner, H. (2014). Three decades of driver assistence systems. *IEEE Intell. Transp. Syst. Mag.* 6, 6–22. doi: 10.1109/MITS.2014.2336271

Clark, A. (2008). *Supersizing the Mind*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195333213.001.0001

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Cole, M., and Engeström, Y. (1993). "A cultural-historical approach to distributed cognition," in *Distributed Cognitions: Psychological and Educational Considerations*, ed G. Salomon (Cambridge, UK: Cambridge University Press), 916–945.

Da Lio, M., Doná, R., Rosati Papini, G. P., and Gurney, K. (2020). Agent architecture for adaptive behaviors in autonomous driving. *IEEE Access* 8, 154906–154923. doi: 10.1109/ACCESS.2020.3007018

Da Lio, M., Doná, R., Rosati Papini, G. P., and Plebe, A. (2022). The biasing of action selection produces emergent human-robot interactions in autonomous driving. *IEEE Robot. Autom. Lett.* 7, 1254–1261. doi: 10.1109/LRA.2021.3136646

Da Lio, M., Mazzalai, A., Gurney, K., and Saroldi, A. (2017). Biologically guided driver modeling: The stop behavior of human car drivers. *IEEE Trans. Intell. Transp. Syst.* 19, 2454–2469. doi: 10.1109/TITS.2017.2751526

Dror, I. E., and Harnad, S. (eds.). (2008). *Cognition Distributed: How Cognitive Technology Extends Our Minds*. Amsterdam: John Benjamins. doi: 10.1075/bct.16

Endsley, M. R. (2017). From here to autonomy: lessons learned from human-automation research. *Hum. Factors* 59, 5–27. doi: 10.1177/0018720816681350

Flemisch, F., Abbink, D., Itoh, M., Pacaux-Lemoine, M.-P., and Weßel, G. (2016). Shared control is the sharp end of cooperation: towards a common framework of joint action, shared control and human machine cooperation. *IFAC-PapersOnLine* 49, 72–77. doi: 10.1016/j.ifacol.2016.10.464

Flemisch, F., Altendorf, E., Canpolat, Y., Weßel, G., Baltzer, M., Lopez, D., et al. (2017). "Uncanny and unsafe valley of assistance and automation: First sketch and application to vehicle automation," in *Advances in Ergonomic Design of Systems, Products and Processes*, ed T. Inagaki (Cham: Springer), 319–334. doi: 10.1007/978-3-662-53305-5_23

Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., and Beller, J. (2012). Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cogn. Technol. Work* 14, 3–18. doi: 10.1007/s10111-011-0191-6

Flemisch, F. O., Adams, C. A., Conway, S. R., Goodrich, K. H., Palmer, M. T., and Schutte, P. C. (2003). *The H-Metaphor as a Guideline for Vehicle Automation and Interaction*. Technical report. NASA.

Fodor, J. (1983). *Modularity of Mind: and Essay on Faculty Psychology*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/4737.001.0001

Gardner, H. (1985). *The Mind's New Science-A History of the Cognitive Revolution*. New York, NY: Basic Books.

Gibson, J. (1986). "The theory of affordances," in *The Ecological Approach to Visual Perception* (Mahwah, NJ: Lawrence Erlbaum Associates), 127–143.

Holzmann, F., Flemisch, F. O., Siegwart, R., and Bubb, H. (2006). *From Aviation Down to Vehicles-Integration of a Motions-Envelope as Safety Technology*. Technical report, SAE Technical Paper. SAE International. doi: 10.4271/2006-01-1958

Hutchins, E. (1995a). *Cognition in the Wild*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1881.001.0001

Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cogn. Sci.* 19, 265–288. doi: 10.1207/s15516709cog1903_1

Jain, A., Del Pero, L., Grimmett, H., and Ondruska, P. (2021). Autonomy 2.0: why is self-driving always 5 years away? *arXiv preprint arXiv:2107.08142*. doi: 10.48550/arXiv.2107.08142

Marcano, M., DIaz, S., Perez, J., and Irigoyen, E. (2020). A review of shared control for automated vehicles: theory and applications. *IEEE Trans. Hum. Mach. Syst.* 50, 475–491. doi: 10.1109/THMS.2020.3017748

Menary, R. (ed.). (2010). *The Extended Mind*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262014038.001.0001

Michalke, T., and Kastner, R. (2011). The attentive co-pilot: towards a proactive biologically-inspired advanced driver assistance system. *IEEE Intell. Transp. Syst. Mag.* 3, 6–23. doi: 10.1109/MITS.2011.941911

Minsky, M. (1986). *The Society of Mind*. New York, NY: Simon and Schuster.

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Newen, A., Bruin, L. D., and Gallagher, S., editors (2018). *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780198735410.001.0001

Plebe, A., Kooij, J. F., Rosati Papini, G. P., and Da Lio, M. (2021). "Occupancy grid mapping with cognitive plausibility for autonomous driving applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Virtual Conference), 2934–2941. doi: 10.1109/ICCVW54120.2021.00328

Poirier, P., and Chicoisne, G. (2008). "A framework for thinking about distributed cognition," in *DrorHarnad:2008* (Amsterdam), 25–43. doi: 10.1075/bct.16.03poi

Rumelhart, D. E., and McClelland, J. L. (eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/5236.001.0001

SAE (2021). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. SAE.

Smart, P. (2017). Extended cognition and the internet-a review of current issues and controversies. *Philos. Technol.* 30, 357–390. doi: 10.1007/s13347-016-0250-2

Vanholme, B., Gruyer, D., Glaser, S., and Mammar, S. (2011). "A legal safety concept for highly automated driving on highways," in *2011 IEEE Intelligent Vehicles Symposium (IV)* (Baden-Baden), 563–570. doi: 10.1109/IVS.2011.5940582

Wang, J., Huang, H., Li, K., and Li, J. (2021). Towards the unified principles for level 5 autonomous vehicles. *Engineering* 7, 1313–1325. doi: 10.1016/j.eng.2020.10.018

Xing, Y., Lv, C., Cao, D., and Hang, P. (2021). Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. *Transp. Res. Part C* 128:103199. doi: 10.1016/j.trc.2021.103199

frontiers | Frontiers in Artificial Intelligence

TYPE Conceptual Analysis
PUBLISHED 17 November 2022
DOI 10.3389/frai.2022.915922

OPEN ACCESS

EDITED BY
Julita Vassileva,
University of Saskatchewan, Canada

REVIEWED BY
Chien-Sing Lee,
Sunway University, Malaysia
Tobias Ley,
Tallinn University, Estonia

*CORRESPONDENCE
Daniel Wolferts
daniel.wolferts@fit.fraunhofer.de

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 08 April 2022
ACCEPTED 31 October 2022
PUBLISHED 17 November 2022

CITATION
Wolferts D, Stein E, Bernards A-K and Reiners R (2022) Differences between remote and analog design thinking through the lens of distributed cognition. *Front. Artif. Intell.* 5:915922. doi: 10.3389/frai.2022.915922

Differences between remote and analog design thinking through the lens of distributed cognition

Daniel Wolferts[1]*, Elisabeth Stein[1], Ann-Kathrin Bernards[2] and René Reiners[1]

[1]Department for Human-Centered Engineering and Design, Fraunhofer Institute for Applied Information Technology, Sankt Augustin, Germany, [2]Department for Data Science and AI, Fraunhofer Institute for Applied Information Technology, Sankt Augustin, Germany

Due to the huge surge in remote work all over the world caused by the COVID-19 pandemic, today's work is largely defined by tools for information exchange as well as new complex problems that must be solved. Design Thinking offers a well-known and established methodological approach for iterative, collaborative and interdisciplinary problem solving. Still, recent circumstances shed a new light on how to facilitate Design Thinking activities in a remote rather than an analog way. Due to Design Thinking's high production of artifacts and its focus on communication and interaction between team members, the theory of Distributed Cognition, specifically the Distributed Cognition for Teamwork (DiCoT) framework, provides an interesting perspective on the recent going-remote of Design Thinking activities. For this, we first highlight differences of analog vs. remote Design Thinking by analyzing corresponding literature from the recent years. Next, we apply the DiCoT framework to those findings, pointing out implications for practical facilitation of Design Thinking activities in an analog and remote setting. Finally, we discuss opportunities through artificial intelligence-based technologies and methods.

KEYWORDS

human-computer interaction (HCI), artificial intelligence (AI), distributed cognition for teamwork, Design Thinking (DT), remote work

1. Introduction

In recent years, due to the COVID-19 pandemic, the world experienced a spike in new digital work and new ways of learning (Brynjolfsson et al., 2020; De' et al., 2020; Feldmann et al., 2021). A lot of professional collaboration between individuals, as well as their interaction with work tools, have become digitized, which affects their work environment and thus results in behavior change within teams. This includes video conferencing tools that have become the go-to mode of communication in team meetings and digital whiteboards or other collaborative software tools that support

creativity and productivity (Unger et al., 2021). With industries shifting their focus from production to more service oriented knowledge work, the need for innovative solutions has been growing constantly (Brown, 2008; Kane et al., 2018). For this purpose many industries have started thinking outside the box and turning to previously unfamiliar disciplines to find new ways of innovative working and problem solving. Design is one of those disciplines and *Design Thinking* (from here on referred to as *DT*) has become a popular framework to facilitate the creation of innovation and to find new solutions to complex, so called wicked problems (Buchanan, 1992). Wicked problems are those type of problems that have, among other attributes, (a) no definitive predefined problem formulation, (b) no stopping rule (i.e., the problem solver can always do better), and (c) no definitive right-or-wrong solution criteria catalog (Kunz and Rittel, 1972). With the ongoing digitization of tools and artifacts, such as collaboration platforms or design tools, these supporting technologies are also becoming "smarter." For instance, Suleri et al. (2019) have introduced an Artificial Intelligence (AI)-powered prototyping tool that lets designers create low-fidelity prototypes and evolves them into mid- to high-fidelity design drafts. Add to this the current trend of going-remote and related discussions of a "post-pandemic workplace" (Kane et al., 2021), we find it necessary to examine the implications of these technological trends on DT practices. The aim of our study is twofold: We first want to highlight and examine differences between analog and remote DT practices. Second, we want to assess the applicability of Distributed Cognition as a guiding theory for researching DT practices in both, analog and remote settings. For this reason, we look at DT and what the going-remote means for DT practices and practitioners. Due to DT's high production of artifacts and its focus on communication and interaction between team members, we use the theory of Distributed Cognition (DCog) as a lens to examine how interactions in the remote differ from interactions in the analog world. Authors like Blandford and Furniss (2005), Webb (2008), or Deshpande et al. (2016) have already looked at Distributed Cognition as a theory to inform research on collaboration in (agile) teams. Our research expands on this notion by examining collaboration in DT, specifically when conducted remotely, and what this means for Design Thinkers' interaction with artifacts, as well as with other individuals. We conclude by alluding to the role of AI in these interactions and highlighting the limitations of this paper as well as future research directions.

2. Background

In this section we present the underlying concepts behind DT and DCog. After that, we introduce our methodological approach to examine the differences between remote and analog DT with respect to DCog.

2.1. Design thinking

DT, despite its growing popularity, is not a clearly defined and universally agreed upon concept. It rather serves as an umbrella term for a diverse conglomerate of *understandings* about human-centered, agile, multi-disciplinary, and creative ways of creating new solutions to existing problems. Scientific literature reveals several attempts to describe attributes common to the different DT approaches.

In its origins, DT has largely been defined by the work of designers (see e.g., Brown 2008), but has now become a multi-disciplinary framework—or as Lindberg et al. (2010, p. 35) put it: "Design thinking understood as a meta-disciplinary methodology loosens the link to design as a profession." Accordingly, DT has increasingly found its way into several research and application domains and practitioners and researchers from different fields of application have started to embrace the "designerly ways of knowing" (Cross, 1982). For example, Kimbell (2011, p. 295) suggests that DT de-politizises managerial practice, in that it helps managers to "shift from choosing between alternatives to helping them generate entirely new concepts." DT has also sparked the interest of psychologists in terms of individual behavior and group dynamics. In this vein, Liedtka (2015) point to DT's potential to reduce cognitive biases in decision making, for example through methodologies that are innate to DT, like perspective-taking, working in teams, or a strong reliance on empirical evidence. Roberts et al. (2016) examine DT's potential for health care, in that it can help health-care professionals to find solutions to complex and overarching problems, like the increase in diabetes and obesity, and help them to bridge the gap between abstract and high-level issues and physicians' day-to-day work. Depiné et al. (2017) describe the integrative nature of DT in their study about Smart Cities. They identified DT's potential not just for developing new technological solutions, but also for integrating citizens needs and concerns as an integral part of the process.

Brenner et al. (2016) understand DT as a triad of *mindset*, *process*, and *toolbox*. With *mindset* they describe a number of guiding principles that Design Thinkers follow, like human-centeredness, applying divergent and convergent thinking, early prototyping, and creating the right environment for creative problem solving. As for *process*, they define an iterative five-step loop of activities regarding problem definition, need-finding and synthesis, ideate (i.e., idea brainstorming), prototyping, and testing. Lastly, they describe a number of tools and methods as the *toolbox* of Design Thinkers, like observation, storytelling, personas, and empathy maps. Its lack of a clear definition, however, can be considered being part of it strengths, because it allows DT "to be the right thing at the right time" (Zimmerman et al., 2007, p. 494). Brown (2008, p. 1) calls DT "a methodology that imbues the full spectrum of innovation activities with a human-centered design ethos." It can be described as a "system of spaces," in which different types of activities take place.

DT models usually follow a multi-stage process, in which activities fall into different categories like exploration, ideation, and materialization. For the purpose of this work, we follow a five step model, for instance as in FIT (2019). The five steps entail: *Empathize*, *define*, *ideate*, *prototype* and *evaluate*. Figure 1 illustrates the iterative DT process. First, *empathize* concerns activities that help the design thinker to build a deep understanding of their target audience and their real-life problems through interviews, desk research, or observation. The gathered information is subsequently structured in the *define* stage through thematic analysis, persona building, or the definition of an actionable problem statements. *Ideation* concerns activities that support the exploration of novel solutions through various brain-storming techniques. Their subsequent incarnation, usually as low-fidelity prototypes that become more sophisticated over time, is carried out in the *prototype* stage. In *evaluate*, design thinkers gather feedback for the created prototypes. As indicated in the figure, DT activities seldom follow each other in a linear fashion. They are rather applied based on their situational necessity.

Examples of the successful application of DT are abundant. In "Creative Confidence," Kelley and Kelley (2013) describe how DT has helped in the design of MRI machines for the pediatric station of a hospital in the USA. Due to children's nervousness and anxiety of the sterile, small, and noisy tubes it can be difficult for radiologists to get a readable image. By methods of observation, interviewing, and iterative prototyping, the design thinkers could develop a redesigned MRI machine and an accompanying room concept that was not met with fear or nervousness by the patients, for instance a pirate ship-themed MRI room. Hehn and Uebernickel (2018) describe how DT can be used for requirements engineering. In their paper, they have analyzed projects from a data base of a Swiss-German consultancy, which were carried out following a DT approach. *Project Falcon*, for instance, was conducted over the course of 20 months by an interdisciplinary team that was comprised of domain experts, designers and business modeling experts to accommodate for feasibility, desirability and viability of the final product. Activities out throughout the project entailed interviews, persona building, mapping out customer journeys, focus groups, prototyping of mock-ups, user tests, and development of the software.

The concept of co-creation is essential to Design Thinking (Plattner et al., 2012). Consequently, a lot of design-thinkers' work happens in collaborative settings (Kress and Schar, 2012). Design happens as a conversation "[. . .] with the problem that is being addressed, with materials and artifacts, with our colleagues and with ourselves" (Sirkin et al., 2012, p. 173). While often not stated explicitly, the *DT workshop* in teams of three to five team members can be found in the literature as the preferred modus operandi of co-creative DT activities (see e.g., Brown, 2008; Levy and Huli, 2019; Schwemmle et al., 2021). We attribute this to at least three factors: One being that a workshop can provide a safe space for the participants which can help spark creativity (Daniel, 2020). The second is that workshops usually yield a specific outcome and high output–something that is highly valued in DT (Mueller-Roterberg, 2018). The third factor that a DT workshop can be found especially in scientific literature is that it is a thankful object of study for researchers. Especially in ethnographic studies, individual and group behavior as well as the artifacts that result from the workshops can be observed and studied quite easily (e.g., in Levy and Huli, 2019). Over the course of this paper, we therefore focus on interactions within DT activities such as DT workshops.

2.2. Theory of distributed cognition

The theory of DCog was developed by Edwin Hutchins in the 1980's and 90's (Hutchins, 1995a). It is an extension to classic Cognition Theory, in that it sees cognitive processes not limited to the individuals' brains, but rather as being distributed in socio-cultural systems. By this token, the processing and the execution of cognitive tasks take place in the interaction and coordination between individuals and their environment, rather than isolated in an individual brain (Zhang and Norman, 1994; Hutchins, 1995a). With this, Hutchins and his colleagues expand the "boundaries of the units of analysis" of cognition research beyond the individual, and toward "the functional relationships of elements that participate together in the process" (Hollan et al., 2000, p. 175). Cognitive processes can hereby be distributed *between members* of a group, between *internal* and *external* representations, and through *time* (Hollan et al., 2000).

In his works, Hutchins studied so called *systems of cooperation*. He identified several artifacts people use to solve complex tasks. In commercial airline flights, for instance, flight crews need to coordinate their tasks and cooperate in order to successfully execute the flight plan. Hutchins propagates that the expertise to fulfill this task resides not only in one individual crew member, but also in the organization of the tools that the crew members use to solve this task. Thus, no single individual can be attributed to being the actual problem solver. Instead, the complex interplay of many actors and artifacts contribute to the successful achievement of the system's goal (Hutchins, 1995b).

To enhance the understanding and analysis of DCog within small teams, Blandford and Furniss (2005) have developed Distributed Cognition for Teamwork (DiCoT) as a method of analysis. They divide it into the three themes: *physical layout*, which focuses on a physical as well as virtual spaces and its objects, *information flow*, which mainly takes the movement and transformation into account and lastly, *artifacts*. As DCog does not focus solely on one individual, but rather "a collection of individuals and artifacts and their relations to each other in a particular work practice" (Rogers and Ellis, 1994, p. 123), the DiCoT framework takes these differences into account. Thus, far DiCoT has mostly been applied in critical systems in a real

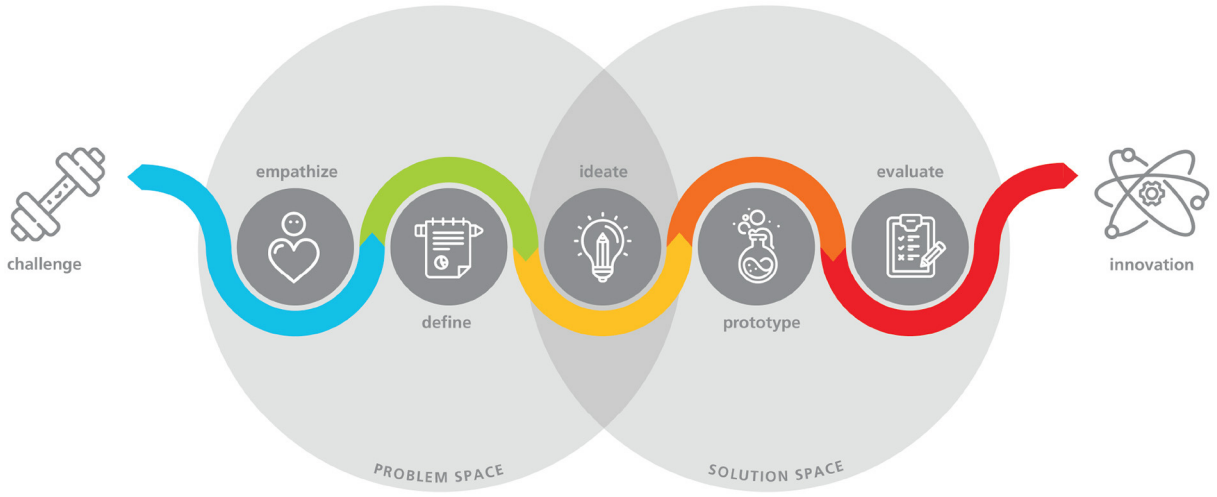

FIGURE 1
Design thinking process by Fraunhofer FIT (FIT, 2019).

environment, e.g., in an emergency medical dispatch (Furniss and Blandford, 2006).

3. Method

First, to develop an understanding of current findings on analog and remote DT and according practices, we examined the relevant literature in the research fields of DT and Remote Facilitation of DT. For this, we applied an approach adapted from Mayring's qualitative content analysis (Mayring, 2021). We started by conducting a Google Scholar search with a subsequent snowballing of the references that could be found in the available literature. We chose Google Scholar as the preferred database, because it covers a wide range of material, including scientific papers, gray literature, technical reports, and conference proceedings. We applied the search terms "Design Thinking," "Remote Design Thinking," and "Distributed Design Thinking." We identified 19 articles of potential relevance, 12 of which were journal articles or conference proceedings, six book chapters and one extended abstract. These documents were then distributed among the authors.

Next, each of the authors highlighted important text segments from the selected papers individually, after which the highlighted segments were collected and their main statement was extracted. For instance, the text segment:

> *"This means that a participant in a workshop can no longer be considered the "victim" of the spatial planning by an architect or interior designer. She rather becomes co-creator of the space through her interaction with spatial elements. If people involved in Design Thinking realize this shift of power and the active role they can take, it allows them to move from accepting space as-is to changing or even preparing a space based on what best suits their requirements."* (Schwemmle et al., 2021, p. 125)

was condensed to *"Participants have agency in actively shaping the space."* In total, 114 of such text segments were highlighted and summarized. Next, each segment was categorized either into *analog* or *remote*, depending on whether they related to analog or remote DT practices. Based on these two categories, we then derived key themes of analog and remote DT practices. With this, we identified the four themes: *Creative Collaboration*, *Space*, *Artifacts*, and *Information Management*. Next, we described those four themes with regard to analog and remote DT practices, the results of which can be found in chapter 4.1.

As we have alluded to earlier, DT usually takes place in small teams: "Team-based working modes are an integral part of Design Thinking. Those teams, especially in corporate environments, are increasingly distributed between locations over the globe" (Wenzel et al., 2016, p. 15). Team members of the design thinking team are therefore affected by the concept Distributed cognition when working on complex tasks in the design thinking process. This applies e.g., for working with artifacts: "Throughout the design-thinking process, the team produces several tangible artifacts: empathy maps, journey maps, storyboards, and wireframes, to name a few." A concept that is known as *representations* in DCog (Hollan et al., 2000; Gibbons, 2016). Hence, in the next step of this paper we aim to deeper understand and learn what DCog might provide to deeper understand the identified themes of remote and analog DT. For this, we drew on the DiCoT framework put forward by Furniss and Blandford (2010), as it is especially

suited to facilitate the application of DCog theory on teamwork settings. We then took the categories that we inferred from the literature review of DT and compared them side by side to the superordinate categories (and their respective principles) we found in Furniss and Blandford (2010). We applied the DiCoT framework as a lens to look at the results documented in the spreadsheet, described possible critical elements that could occur when practicing remote DT. The results are described in chapter 4.2. From this, we lastly derived implications for practice when conducting DT activities remotely as opposed to analog.

4. Results

In this section we first describe the themes we found when analyzing DT literature and present differences between remote and analog DT practices for each theme. Further, we continue by applying the theory of Distributed Cognition and more specifically the DiCoT framework, to understand and frame these differences from a DCog perspective.

4.1. Results of literature review on design thinking

The analysis of the literature yielded four themes of analog and remote DT practices, namely *creative collaboration*, *space*, *artifacts*, and *information management*. The following sections describe the findings by juxtaposing practices applied in analogous settings with practices from digital settings. Table 1 summarizes the key findings. Especially the past 2 years have proven to be important as remote work experienced a huge surge due to the COVID-19 pandemic.

4.1.1. Creative collaboration

4.1.1.1. Analog

Creative collaboration between team members is at the heart of DT. Creative collaboration can be influenced by intentionally composing the DT teams with members from different disciplines and cultural backgrounds. Due to its strong emphasis on team-based learning, DT helps to "extend mono-disciplinary rationales by offering a flexible meta-rationale, which counters the restriction of admissible questions or analytical schemes typical of mono-disciplinary thinking" (Lindberg et al., 2010, p. 35).

Collaboration and space are tightly connected with each other. According to Schwemmle et al. (2021), collaborative teams 'create' the space around them, not just by replacing and altering the elements in the space. Rather, the team members construe meaning to the space through interaction with and perception of the space. In this vein, the interaction of individual team members with the space is perceived, either directly or indirectly, by other team members, which influences the behavior and therefore the interaction of those other team members with the space. This type of reciprocal relationship between individuals, space, and teams can lead the team to perceive a space as 'their space' or 'their home', which might provoke a feeling of territoriality (Brown et al., 2005) and, in extension, a sentiment of psychological ownership for the space (as in Dawkins et al., 2017). By distributing and placing certain elements in the space (e.g., chairs, tables, whiteboards, etc.), DT facilitators are able to determine the character of the collaboration between the team members. That way, a team can be prompted to work in a self-organized manner, instead of a hierarchical task distribution. In addition, the creation of a social (sub-)space might foster a positive atmosphere and provide them with a safe-space where they can nurture their personal relationships, rather than work on a specific task (Schwemmle et al., 2021). This phenomenon not just holds true for the concept of space, but also for ideas (Elsbach and Flynn, 2013) where individuals feel like they "own" ideas and show competitive of defensive behavior as part of the "possessive self" (De Dreu and Van Knippenberg, 2005).

4.1.1.2. Remote

Creative collaboration in remote settings is largely influenced by tools that mediate communication. Luther and Bruckman (2008, p. 343) define online creative collaboration as "[…] comprising two key properties. First, people communicate and meet each other chiefly *via* computer-mediated communication. Second, they do so with the purpose of working together to create new artifacts." Donaldson et al. (2021) describe several benefits of remote collaboration over an analogous setting, like lower costs due to decreased traveling and used up material, potentially higher retention rates due to less effort for the team members to attend, and better scalability of workshops due to the lack of spacial restrictions like room sizes. However, according to Vallis and Redmond (2021), there is a persisting inaccessibility of digital whiteboards and drawing tools to large mainstream cohorts, which leads to the exclusion of certain populations from the design process.

4.1.2. Space

4.1.2.1. Analog

For DT practices in an analog setting, it is important to look at the role a physical space plays for the applied practices. The physical space in which a DT workshop takes place is filled with elements such as seating opportunities, tables, flip charts or whiteboards and other physical elements to work with (Schwemmle et al., 2021). Especially when working together collaboratively on certain tasks people gather around whiteboards or create sub-spaces within a room with portable walls. But not only is space important because of its physical elements "such as its floor plan, distances, or atmospheric cues as perceived by the user (perceived constructed space)" but it is also defined through what people bring with them "such as the

TABLE 1 Key results of literature review for remote and analog Design Thinking.

Theme	Remote	Analog
Creative Collaboration	• The quality of creative collaboration is highly influenced by the attributes of the tools that mediate it. • Low costs due to less traveling and less used-up material. • Potentially higher retention rates due to less effort to participate. • Improved scalability of workshops. • Online collaboration tools like digital whiteboards are still inaccessible for large, mainstream cohorts.	• Collaboration extends mono-disciplinary rationales. • It is highly influenced by the space that surrounds people. • Creative collaboration can foster a feeling of ownership and defensive behavior with respect to ideas or the space.
Space	• Digital tool is essential part of the team's success. • They have no physical boundaries. • Extended "space" by enable external cues from the internet. • Asynchronous and synchronous work possible • Meetings have to be scheduled, and are. cannot happen spontaneous.	• Physical room with physical elements (chairs, whiteboards etc.) allows for spontaneous creation of sub-spaces. • People "create" the room in a reciprocal interaction. • Physical space can become a 'home'.
Artifacts	• Digital artifacts lack "materiality." • They lose their physical restrictions and their functional qualities with respect to their material attributes. • Digital artifacts can become "smart" through AI (e.g., in *situated agents*).	• Designers transform ideas into "tangible representatives," i.e., artifacts. • They facilitate communication and collaboration internally and externally. • Artifacts can guide reflective behavior.
Information Management	• Easy to access information through the internet or knowledge management tools. • Asynchronous work is good for information gathering. • Synchronous work enables discussing the value of the gathered information.	• Mostly synchronous work. • Information needs to explicitly made accessible and reproducible. • All information at hand/in the room

individual's perception, its experiences and resulting behaviors (reflected constructed space)" (Schwemmle et al., 2021, p.125). Because space is constantly changing throughout the process of iterative work, it must be understood on a behavioral level as well (Brown, 2008). People in a designated space are not merely caged in it, but they rather "create" space by interacting with the physical elements in it and assigning meaning to these elements or their arrangement. The physical space influences people but can also be influenced by them (Sirkin, 2011; Schwemmle et al., 2021), thus forming a reciprocal relationship as touched upon earlier. When it comes to the atmospheric perception of space, for DT practices it is necessary to let people feel like they have a *home*. This helps to foster co-creation and inspiration as people feel that a certain space is "their own" which "creates safety for a team, allows identification and fosters well-being" (Schwemmle et al., 2021, p.133).

4.1.2.2. Remote

In remote DT a variety of digital tools can be used that set up and guide the scenery of DT activities. They provide a digital space where a physical space is not available. This requires having a tool to communicate during the workshop and a tool to collaborate in a virtual space (Sirkin, 2011). Wenzel et al. (2016, p. 16) emphasize that "the digital tool is not only a plain functional instrument. In fact, with its usability and acceptance it is a relevant factor for the teams' success. Thus, making the tool an important player among the members of a team requires its interplay with the team's working situation." In consequence, a virtual space where teams can coordinate and carry out their work and collaborate on specific tasks and design a solution is an essential part of the remote DT process. Additionally, a virtual space has no physical boundaries. Especially when it comes to innovative thinking and the creation of ideas, extending ones space beyond the immediate surroundings using digital tools might offer a wider perspective and inspiration (Unger et al., 2021). Team members can access external cues such as images or information *via* search engines and extend their knowledge (Vallis and Redmond, 2021).

Along with the workshop design that might integrate single person work time or might even schedule the process within

an extended period of time, a virtual space holds the chance to work on a personal scale and speed and at the same time synchronize the teams' progress, which is a challenge to master (Yarmand et al., 2021). Asynchronous work however, might lower hesitations in creative thinking such as the fear of speaking up, etc. Nonetheless, a virtual space is a challenge when it comes to spontaneous collaboration and might even hinder team members to get help in the moment they really need it (Sirkin, 2011).

4.1.3. Artifacts

4.1.3.1. Analog

The creation of and interaction with artifacts is crucial to DT: "Design-as-practice cannot conceive of designing (the verb) without the artifacts that are created and used by the bodies and minds of people doing designing" (Kimbell, 2012, p. 135). Designers transform ideas into tangible representatives (i.e., artifacts), which not only facilitate communication and collaboration within the team, but also with external stakeholders and help the designers to stay in touch with the problem-relevant environment (Lindberg et al., 2010). For Jung and Stolterman (2010, p. 153), "design can be considered as a process of creating meaning with proper materials based on exploratory practice with them," thus highlighting the importance of physicality and tangibility of artifacts in DT. Ghajargar and Wiberg (2018, p. 5) describe the potential for artifacts to guide reflective behavior, with "'reflection' referring to the action of reflecting on information provided, and being informed about the consequences of an action or behavior and to create puzzling and surprising effects". As physical representations of an idea, artifacts as prototypes help designers not just to explore their solutions, but also to communicate these solutions with the outside world. These types of artifacts, however, do not only have a profound influence on the actors in a design process—they also play an important role in research. In design research and HCI the *research through design* approach lets researchers examine problem spaces and create solutions "through the construction of artifacts" by applying methods informed by the work of designers (Zimmerman and Forlizzi, 2008, p. 42).

4.1.3.2. Remote

Due to their non-physicality and non-tangibility, digital artifacts play somewhat of a different role in DT compared to their analog counterparts. According to Balters et al. (2021, p. 10f), "[t]he use of artifacts (analog or digital) affects practically every facet of the DT methodology and practice. The absence or curtailment of artifact usage and accessibility combined with the absence of face to-face interaction together severely changes the quality of interaction and outcome between people during an innovation event." On the one hand, as digitization progresses, artifacts (such as prototypes or work tools) lose their materiality, or become *materials without qualities*, as Löwgren and Stolterman (2004) put it. Thus, artifacts lose their physical restrictions and their functional qualities with respect to their material attributes. This becomes especially interesting if we consider spatial artifacts as tools for the design process, like the properties of a room or the furniture in it. Examining this loss of material quality becomes increasingly important, as more and more collaborative design happens digitally and artifacts become digitized. On the other hand, artifacts gain certain attributes as well, for instance, as tools are becoming increasingly smart. AI-supported tools, so called *situated, reactive* or *behavioral agents* are context aware, proactive (i.e., can act autonomously), preemptive (i.e., help humans to prevent errors) and interactive (Ghajargar and Wiberg, 2018).

4.1.4. Information management

4.1.4.1. Analog

Information gathering and processing in DT activities are important factors to successfully running for example DT workshops. This counts for existing knowledge as input for and output of these activities as well. Tracing information about design decisions throughout all phases of the DT process is therefore key for iterative work. If information gets lost "the evaluation of ideas is often restricted to the prototype alone and cannot be navigated back to the original source of the design decisions" (Gabrysiak et al., 2011, p. 221). Analog group work is mostly synchronous, all the necessary information is either contributed by the DT facilitator or by the team members themselves (Schwemmle et al., 2021). It is therefore important that all information in the room of the DT workshop is accessible as well as reproducible at any time (Gabrysiak et al., 2011). When communicating the result of a DT activity, a tested prototype for e.g., a product, the engineering team has to be aware of design decisions made so they can take them into account in their work. Information not transferred might lead to the end product not being in line with the envisioned idea of the DT Team.

4.1.4.2. Remote

Remotely working in the DT phases on the one hand makes it easy to access information, e.g., by searching for additional information online and by using the landscape of free digital tools that support design processes. This might help bringing everyone on board, "when information has to be shared among all participants–for instance, during the welcome phase of a workshop, introducing the challenge or giving interim inputs, and, finally, to present results to all participants and maybe even to an external audience" (Schwemmle et al., 2021, p. 129). On the other hand a challenge in a remote setting is to develop a common knowledge base. Here, working asynchronously for getting more information but synchronously for discussing the value of it might be necessary.

4.2. Application of DiCoT on DT

After having identified four themes in the DT literature that highlight the differences between remote and analog DT practices, we aim at understanding the implications of the ongoing digitization on those practices. For this purpose, we use DiCoT as a framework to look at our findings from a DCog's perspective. DiCoT has been applied as an analytical framework for collaborative work settings before (e.g., in Hussain and Weibel 2016). Given the highly collaborative nature of DT, we consider this framework particularly suited for helping us uncover why the differences between analog and remote DT occur. Blandford and Furniss (2005) describe their themes *physical layout*, *information flow* and *artifacts*. Each of their themes is further divided into different *principles*, which we will consider in the context of DT in the following.

4.2.1. Physical layout

The *physical layout* includes all the environmental aspects influencing the performance of the cognitive system. Those environmental aspects may refer to auditive, visual, or tactile stimuli, which shape the perceptions of humans and thus have a direct impact on their computational capabilities.

When it comes to remote DT practices *space and cognition*, which are one of the principles of the physical layout, differ from an analog setting. In the analog world, a table can be moved and stacked with papers, thus help people to reduce complexity and make choices. In the remote setting space is effectively infinite and not limited to borders e.g., of a table. Artifacts that are out of sight cease to exist, which might increase the complexity of collaboration. Also, digital representations of information are not easily perceived as natural (*naturalness principle*). Even though digital tools try to copy the real world whenever possible, e.g., a sticky note in real life and its representation in an online tool are much alike, it is especially the interaction with digital representations which differs from the interaction with analog ones. This might provide more effort for mental transformations to make use of those representations.

Furthermore, tools for remote collaborative work can provide assistance when prioritizing content or tasks (e.g., the "bring everyone to me"-function, or the mini-map in the collaborative online tool Miro), which in the DiCot framework can be found in the *perceptual principle*.

The principle of *situation awareness* describes that people need to be informed about details of a situation such as what is planned and what is going on. Blandford and Furniss (2005, p. 29) even state: "The quality of this situation awareness can be influenced by how accessible the work of the team is." In the analog context the proximity toward the team is important, which means one can observe or even overhear what is happening. In a remote setting there is a lack of proximity which might lead to less situation awareness. The digital landscape of supporting tools for remote work also provide the opportunity to get a better overview on what other teams work on. Especially for interdisciplinary teams working globally this might create perceived proximity to other teams. This also holds for the *horizon of observation* which provides the team member with an overview of everything that can be seen and heard. Even though being present in one physical space helps by focusing on the environment, a remote access to all information and digital possibilities of documentation such as video recording or chatting might broaden the horizon of observation as it can be accessed any time and from everywhere. However, this requires a moderation that focuses on a common understanding of the teams' horizon of observation. Being aware of ones surroundings also includes the *arrangement of equipment*, which is key to processing information. Here, an analog setting is limited to the space and equipment at hand, which might be of advantage when focusing on reducing complexity of problems. In a remote setting, space and access to information is not limited due to a broad range of digital tools. On the one hand, this might help to understand complex problems and get inspired more thoroughly, but on the other hand could lead to an overload of information that cannot be processed anymore.

The last principle as part of the physical layout is *subtle bodily supports*, which mostly comes into effect in analog DT so far, e.g., when team members can point at things and speak with their body. This is limited in a remote setting as it is not the finger as part of the body pointing on things, but its digital representation, e.g., a cursor (Sirkin, 2011).

4.2.2. Information flow

Information flow revolves around the interaction of entities within the cognitive system. This may include the communication between team members and the transformation of information or tools that facilitate the information flow. Looking at a remote DT workshop representations (i.e., physical realizations of artifacts) are different from their analog counterparts. This has consequences for the information management and therefore flow of information.

First, the principle of *information movement* seems to be present differently within the analog than the remote setting. Information in the analog setting can be, for example, "passing physical artifacts; text; graphical representation; verbal; facial expression; telephone; electronic mail; and alarms" (Blandford and Furniss, 2005, p. 32). In the remote setting, information is mostly provided in a two-dimensional way, e.g., an emoji on an online whiteboard. Both, analog and remote DT activities support information movement. Still, being online might put a different kind of speed and complexity on information movement, as much information can be accessed quickly and with low effort, for instance because it easier to copy and paste information like pieces of text, from A to B.

The principle *transformation of information* describes changes in the representation of information. One example is filtering information, which in the DT process is an essential part. Several artifacts, like sticky notes, are gathered, sifted and structured and thus distilled to one key aspect which is written on a new sticky note (so called *clustering* or *thematic analysis*). The transformation of information can therefore be realized in both, the analog and remote DT process. Some digital tools for remote work even feature tracing prior steps within the process and therefore allow asynchronous collaboration.

An important activity of DT is to channel all information that is necessary to develop innovations, whether it is in a physical or a digital space, so that every team member has access to it at any time from anywhere. The principle of *information hubs* describes that information from different sources are brought and processed together. Especially when a team makes decisions an information hub helps to define further steps. Elements in a physical room (e.g., whiteboards or flip charts) as well as in a digital room (e.g., online whiteboards) can support this decision-making process by displaying all information in one place and therefore promote effective work. However, the distribution of information may differ in its extent and the ways it is provided. When a lot of information is shared at the same time, the principle of *buffering* applies, which aims at withholding information until it can be introduced without the risk of errors or conflicts with ongoing activities.

4.2.3. Artifacts

The third topic centers around artifacts and how their design enhances cognition of the individuals within the system. This includes the layout of an artifact, for instance the form of a sticky note, as well as the physical movement of it, e.g., hanging a sticky note on a chart or moving it. The principle of *mediating artifacts*, for instance, describes those types of artifacts that help the team to complete the task. In remote DT, digital artifacts have different attributes as compared to haptic ones. For example, a lot of people's interactions with artifacts relate back to their positioning in their immediate environment. That means that artifacts can be used to create *scaffolding*—the second principle of artifacts—for example by placing a marker where a task was left and is to be picked up again in the future. In digital spaces, however, Design Thinkers might lose the perception of an artifact's location due to the space's lack of physical boundaries. Also, as alluded to earlier, materiality plays a key role in people's interaction with those artifacts. The absence of the tangibility of a sticky note, for instance, could make it harder for team members to use these artifacts to their advantage (like passing a sticky note to another team member). Artifacts also serve a representational function in that they create *goal parity*—the third principle of artifacts—between the actors involved. Prototypes, for instance, mediate between the contemporary and a desired future state and communicate a DT teams vision and thought process to people external to the team. Digital prototypes, however, do not convey the same type of experience as haptic ones do. Practitioners should therefore pay close attention to the way they create and use artifacts. Given that digital prototypes are for now impossible to touch, smell, or taste, practitioners have to rely on visual and auditive clues for their target audience to represent the desired future state. They should therefore rely on methods like storytelling or scribbling. Additionally, they should pay attention to the arrangement of artifacts on the digital spaces that they are working on. If too many digital artifacts occupy the digital space and get pushed outward, or if too many digital spaces are created, people might lose awareness about their existence.

5. Discussion

5.1. Implications for practitioners

After having elaborated on the perspective of DCog analog and remote DT practices through the lens of the DiCoT framework, we now address specific implications for DT practitioners when applying analog and remote DT. As the ongoing digitization requests applying both, analog and remote DT, differences and how they can be used efficiently have to be considered when designing DT workshops.

As mentioned above, *physical layouts* in remote settings differ from analogs ones. Space is virtually infinite, which can lead to a feeling of being lost or overwhelmed. Practitioners should use tools that allow them to structure the space into working areas and assign themes to these working areas. This could help participants of a DT workshop to better orient and reduce the complexity of the virtual space. Functions like "Bring to me," where the participants' view on a digital whiteboard is pulled to the facilitator, can also help with streamlining attention. In order to create a feeling of naturalness, practitioners should use tools that allow them to create items that represent a natural form, like a digital sticky note. They should also point out their natural representation when they introduce the tools to inexperienced workshop participants. This can give them a cognitive aid and make it more effortless for them to use these items and to feel at ease while using them. The size of the items (e.g., a sticky note) used in the DT process can also be utilized by Design Thinkers. Increasing the size of an item may indicate that it is of higher importance than others. The same holds true for the color or the shape of an item. However, in some situations this could be counterproductive. In brainstormings, for instance, all ideas should initially have equal weight and importance. If certain ideas are displayed on a larger sticky note, this could lead to a selective perception (Pronin, 2007) in the further progress. In general, situational awareness is harder to convey in remote settings. Hence, practitioners should be much more descriptive in their language and explicate most

of what they are currently doing verbally. When conducting workshops with inexperienced Design Thinkers, this needs to be pointed out regularly, as it creates transparency and awareness for the other team members. Additionally, enabling participants to independently move through spaces or rooms (e.g., break-out rooms) promotes autonomy. Generally, tools like digital whiteboards or video communication should be kept as simple as possible, as switching between tools demands cognitive resources from the participants (Skulmowski and Xu, 2022).

To ensure good *information flow*, practitioners should explicate verbally when they move a digital item from one point to the other. Team members' attention can be focused on a different part of the working area and they might not even be aware, that an item was moved. Generally, for team members who are not proficient with tools like digital whiteboards, moving items might not be an easy task. Warm-up games in the beginning of a workshop could help to practice this. A big advantage of digital items is that they can be copied and pasted virtually without effort–other than analogous ones. By copying and pasting created items to other sections of the working area, rather than moving them, the entire process becomes better traceable. Also, by clustering certain information together in information hubs and giving these hubs prominent thematic captions, it can help practitioners to find information quicker and easier. One of these information hubs could be an *Idea Parking Lot*, where team members can "park" their spontaneous ideas for later reuse. Creating spaces where information can be "parked" might help the team to reduce cognitive capacities and therefore focus on the information at hand. Yet, information is not lost but can be introduced at a later stage.

Artifacts, as the tools that help DT practitioners to generate and convey their ideas, can also be used to create structure. Some tools are better developed to create ideas, while others might be better suited for prototyping. Generally, practitioners should think about the purpose at hand prior to the DT activity and then chose a fitting tool. Also, artifacts can create structure, in that they can indicate to the practitioners where in the DT process they currently located, for instance through a Kanban board or other visual cues. Additionally, to make up for the lack of materiality of artifacts like prototypes, practitioners could break through the two-dimensionality of virtual artifacts, for instance by printing out prototypes of a web-page or by providing team members with do-it-yourself kits like Lego Serious-Play, in order for them to rebuild the prototype in real-life.

5.2. Future outlook and AI technologies

Finally, as practitioners in the field of computer science we deem technologies like AI to potentially contribute to the future of DT practices. In the following we therefore want to offer our thoughts what AI might hold for DT practices in the future.

Especially remote DT activities hold much potential for incorporating AI to enhance workshop and learning experiences. For instance, Eve, a software tool that helps designers to create low-fidelity prototypes and transition them into mid-fidelity to high-fidelity prototypes *via* machine learning was presented by Suleri et al. (2019). Such a tool can help design thinkers, not just to learn how to prototype faster and easier, but also to enhance cognition by distributing difficult tasks to the software tool. Another possibility, provided by the use of AI within the physical layout, lies within the arrangement of equipment. Being aware of ones surroundings and the equipment at hand might be supported by AI. For example, an AI-based companion could help a facilitator to raise awareness of available artifacts and could further accompany creative processes by using those artifacts (Verganti et al., 2020). This could also trigger behavior change, as facilitators with lower experience have a higher level of guidance, thus can implement new methods more easily.

For the *information flow*, AI could support the transformation of information, by offering intelligent clustering or filtering (Verganti et al., 2020). This means, the decision, whether an information is important now or if it can be hold up until later, called buffering, which currently has to be made by a human intelligence could also be assisted by AI. In addition, AI could be supportive in making those choices for example with recommender systems. Further, AI algorithms could also highlight the most important information to allow information hubs.

Artifacts can already be supported by AI, e.g., with the "stickies capture" in Miro (smart text recognition), where analog sticky notes are automatically recognized and digitized. Furthermore, the creation of artifacts could also be supported by such AI powered tools, e.g., algorithms that could be included in brainstorming activities, or while gathering information on target groups. This could support facilitators, as well as team members of DT activities.

Participants of DT activities need to be able to trace where, e.g., information for personas are coming from, which might help to minimize the risk of research biases. It is therefore essential that algorithms, especially those that support decision making, are transparent and explainable (Explainable AI; XAI) (Gunning et al., 2019).

5.3. Limitations

This work has been conducted from the perspective of DT facilitators. For future research, the perspective of team members of DT activities such as workshops should be taken into account to broaden the significance of the results for the respective target groups. This paper aims at providing a DT facilitators view on the changing environment, that has been caused by the COVID-19 pandemic, rather than empirical insights.

Thus, collecting primary data could yield further insights to practitioner guidelines and improve the overall experience of team members and facilitators in creative work. This may also foster insights into team members' appropriation of digital tools in DT activities and how this might influence their practice.

Due to the COVID-19 pandemic, the demand for digital tools has increased, as many companies adopted hybrid working modes and therefore needed more and more sophisticated tools to facilitate this change. This has accelerated the development and improvement of various digital tools for remote collaboration, like digital whiteboards or video conferencing tools. Due to the high and increasing demand, these tools grow rapidly in functionality, with providers adding more and more features to their products, rendering remote collaboration easier in some aspects (e.g., asynchronous work) and more complex in others (e.g., efficient communication). As this is a continuous process, publications differ in the functionality in the respective tool that has been used, which may have an impact on the user experience and feasibility of workshops for practitioners. As the pandemic forced many practitioners to quickly switch to a remote environment on a day-to-day basis, recent reflective insights on findings for remote DT might not have been published yet at the time of this paper.

This work aimed at looking at the DT process as a whole and not divide it into its five phases. For future research, it may be of interest to examine how the different phases are executed in analog and remote settings and where they differ most with regards to interaction and distributed work. This would be an even closer look into DT activities and might allow for more precise practitioner guidelines. Even though there is a lot potential for the integration of AI into DT practices, it also holds risks as well. These may include increasing the cognitive load (Skulmowski and Xu, 2022) and no sufficient technical competency by team members and facilitators.

6. Conclusion

We have pointed to the potential influence that digitization and the development toward increased remote-based work might have on DT practices. We have analyzed scientific literature from the relevant research areas and identified four themes in which remote DT practices might differ from analog ones. We have used the theory of DCog and the DiCoT framework according to Blandford and Furniss (2005) to put our findings in perspective and collect concrete notes for DT practitioners when conducting remote DT workshops. Interactions with digital tools have had an increased importance in many workplaces, which allows us to draw the connection to the theory of DCog. Therefore, a new light has been shed on the relevance of corresponding theories and concepts. This allows future research on remote work settings to take DCog closer into account to evaluate digital tools and interactions with those.

Author contributions

With his expertise in Design Thinking and HCI, DW contributed substantially to the introduction and the literature review of Design Thinking literature, and contributed parts of the section on Distributed Cognition. With her expertise in Human-Centered Design and Design Thinking, ES contributed substantially to the literature review of Design Thinking literature, designed and drafted (Table 1), and contributed to the limitations section. With her expertise in psychology and computer-supported collaborative learning, A-KB contributed substantially to the literature review of Distributed Cognition, and the conclusion and limitations sections. With his Ph.D. in computer science and his experience as the head of department for Human-Centered Engineering and Design at Fraunhofer FIT, RR contributed to the conclusion and future directions of the paper. All authors contributed to editorial tasks in all chapters.

Funding

The research leading to these results has received funding from the European HORIZON 2020 program under grant agreement no 857202, the DEMETER project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Balters, S., Baker, J. M., Hawthorne, G., and Reiss, A. L. (2021). "Inter-brain synchrony and innovation in a zoom world using analog and digital manipulatives," in *Design Thinking Research* (Cham: Springer), 9–32.

Blandford, A., and Furniss, D. (2005). "DiCoT: a methodology for applying distributed cognition to the design of teamworking systems," in *International Workshop on Design, Specification, and Verification of Interactive Systems* (Berlin, Heidelberg: Springer), 26–38.

Brenner, W., Uebernickel, F., and Abrell, T. (2016). "Design thinking as mindset, process, and toolbox," in *Design Thinking for Innovation* (Cham: Springer), 3–21.

Brown, G., Lawrence, T. B., and Robinson, S. L. (2005). Territoriality in organizations. *Acad. Manag. Rev.* 30, 577–594. doi: 10.5465/amr.2005.17293710

Brown, T. (2008). Design thinking. *Harv. Bus. Rev.* 86, 84.

Brynjolfsson, E., Horton, J. J., Ozimek, A., Rock, D., Sharma, G., and TuYe, H.-Y. (2020). *COVID-19 and remote work: an early look at us data*. Technical report, National Bureau of Economic Research.

Buchanan, R. (1992). Wicked problems in design thinking. *Design Issues* 8, 5–21. doi: 10.2307/1511637

Cross, N. (1982). Designerly ways of knowing. *Design Stud.* 3, 221–227. doi: 10.1016/0142-694X(82)90040-0

Daniel, G. R. (2020). Safe spaces for enabling the creative process in classrooms. *Aust. J. Teach. Educ.* 45, 41–57. doi: 10.14221/ajte.2020v45n8.3

Dawkins, S., Tian, A. W., Newman, A., and Martin, A. (2017). Psychological ownership: a review and research agenda. *J. Organ. Behav.* 38, 163–183. doi: 10.1002/job.2057

De Dreu, C. K., and Van Knippenberg, D. (2005). The possessive self as a barrier to conflict resolution: effects of mere ownership, process accountability, and self-concept clarity on competitive cognitions and behavior. *J. Pers. Soc. Psychol.* 89, 345. doi: 10.1037/0022-3514.89.3.345

De', R., Pandey, N., and Pal, A. (2020). Impact of digital surge during COVID-19 pandemic: a viewpoint on research and practice. *Int. J. Inf. Manag.* 55, 102171. doi: 10.1016/j.ijinfomgt.2020.102171

Depiné, Á., de Azevedo, I. S. C., Santos, V. C., and Eleutheriou, C. S. T. (2017). "Smart cities and design thinking: sustainable development from the citizen's perspective," in *Proceedings of the February 2017 Conference: IV Regional Planning Conference* (Aveiro), 23–24.

Deshpande, A., Sharp, H., Barroca, L., and Gregory, P. (2016). "Remote working and collaboration in agile teams track: 15. managing is projects and is development remote working and collaboration in agile teams," in *International Conference on Information Systems* (Dublin, Ireland).

Donaldson, J. P., Choi, D., and Layne, J. (2021). "Online design thinking faculty development workshops: a design-based research study," in *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021* (Bochum, Germany: International Society of the Learning Sciences), 209–212.

Elsbach, K. D., and Flynn, F. J. (2013). Creative collaboration and the self-concept: a study of toy designers. *J. Manag. Stud.* 50, 515–544. doi: 10.1111/joms.12024

Feldmann, A., Gasser, O., Lichtblau, F., Pujol, E., Poese, I., Dietzel, C., et al. (2021). "Implications of the COVID-19 pandemic on the internet traffic," in *Broadband Coverage in Germany; 15th ITG-Symposium* (Berlin, Offenbach), 1–5.

FIT. (2019). *Design Thinking with Fraunhofer*. Fraunhofer FIT - Design Thinking Factory. Available online at: https://www.design-thinking-factory.fit.fraunhofer.de/en/design-thinking.html

Furniss, D., and Blandford, A. (2006). Understanding emergency medical dispatch in terms of distributed cognition: a case study. *Ergonomics* 49, 1174–1203. doi: 10.1080/00140130600612663

Furniss, D., and Blandford, A. (2010). "DiCoT modeling: from analysis to design," in *Proceedings of CHI Workshop Bridging the Gap: Moving From Contextual Analysis to Design* (Atlanta, GA), 10–15.

Gabrysiak, G., Giese, H., and Seibel, A. (2011). "Towards next generation design thinking: scenario-based prototyping for designing complex software systems with multiple users," in *Design Thinking* (Berlin, Heidelberg: Springer), 219–236.

Ghajargar, M., and Wiberg, M. (2018). Thinking with interactive artifacts: reflection as a concept in design outcomes. *Design Issues* 34, 48–63. doi: 10.1162/DESI_a_00485

Gibbons, S. (2016, September 18). Design Thinking Builds Strong Teams. *Nielsen Norman Group*. Available online at: https://www.nngroup.com/articles/design-thinking-team-building/

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). Xai–explainable artificial intelligence. *Sci. Rob.* 4, eaay7120. doi: 10.1126/scirobotics.aay7120

Hehn, J., and Uebernickel, F. (2018). "The use of design thinking for requirements engineering: an ongoing case study in the field of innovative software-intensive systems," in *2018 IEEE 26th International Requirements Engineering Conference (RE)* (Banff, AB: IEEE), 400–405.

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Hum. Interact.* 7, 174–196. doi: 10.1145/353485.353487

Hussain, M., and Weibel, N. (2016). "Can dicot improve infection control? a distributed cognition study of information flow in intensive care," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA), 2126–2133.

Hutchins, E. (1995a). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cogn. Sci.* 19, 265–288. doi: 10.1207/s15516709cog1903_1

Jung, H., and Stolterman, E. (2010). "Material probe: exploring materiality of digital artifacts," in *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction* (Funchal, Portugal), 153–156.

Kane, G. C., Nanda, R., Phillips, A., and Copulsky, J. (2021). Redesigning the post-pandemic workplace. *MIT Sloan Manag. Rev.* 62, 12–14.

Kane, G. C., Palmer, D., Phillips, A.-N., Kiron, D., and Buckley, N. (2018). Coming of age digitally. *MIT Sloan Manag. Rev. Deloitte Insights* 59, 1–10.

Kelley, T., and Kelley, D. (2013). *Creative Confidence: Unleashing the Creative Potential Within Us All, 1st Edn*. Crown Business.

Kimbell, L. (2011). Rethinking design thinking: part i. *Design Cult.* 3, 285–306. doi: 10.2752/175470811X13071166525216

Kimbell, L. (2012). Rethinking design thinking: part ii. *Design Cult.* 4, 129–148. doi: 10.2752/175470812X13281948975413

Kress, G. L., and Schar, M. (2012). "Teamology-the art and science of design team formation," in *Design Thinking Research* (Berlin, Heidelberg: Springer), 189–209.

Kunz, W., and Rittel, H. W. (1972). Information science: on the structure of its problems. *Inf. Storage Retrieval* 8, 95–98. doi: 10.1016/0020-0271(72)90011-3

Levy, M., and Huli, C. (2019). "Design thinking in a nutshell for eliciting requirements of a business process: a case study of a design thinking workshop," in *2019 IEEE 27th international requirements engineering conference (RE)* (Jeju: IEEE), 351–356.

Liedtka, J. (2015). Perspective: Linking design thinking with innovation outcomes through cognitive bias reduction. *J. Product Innovat. Manag.* 32, 925–938. doi: 10.1111/jpim.12163

Lindberg, T., Noweski, C., and Meinel, C. (2010). Evolving discourses on design thinking: how design cognition inspires meta-disciplinary creative collaboration. *Technoetic Arts* 8, 1. doi: 10.1386/tear.8.1.31/1

Löwgren, J., and Stolterman, E. (2004). *Thoughtful Interaction Design: A Design Perspective on Information Technology*. Cambridge, MA: MIT Press.

Luther, K., and Bruckman, A. (2008). "Leadership in online creative collaboration," in *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA), 343–352.

Mayring, P. (2021). *Qualitative Content Analysis: A Step-by-Step Guide*. London: Sage.

Mueller-Roterberg, C. (2018). *Handbook of Design Thinking*. Mülheim: Hochschule Ruhr West.

Plattner, H., Meinel, C., and Leifer, L. (2012). *Design Thinking Research*. Cham: Springer.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends Cogn. Sci.* 11, 37–43. doi: 10.1016/j.tics.2006.11.001

Roberts, J. P., Fisher, T. R., Trowbridge, M. J., and Bent, C. (2016). A design thinking framework for healthcare management and innovation. *Healthcare* 4, 11–14. doi: 10.1016/j.hjdsi.2015.12.002

Rogers, Y., and Ellis, J. (1994). Distributed cognition: an alternative framework for analysing and explaining collaborative working. *J. Inf. Technol.* 9, 119–128. doi: 10.1177/026839629400900203

Schwemmle, M., Nicolai, C., and Weinberg, U. (2021). "Using 'Space' in design thinking: concepts, tools and insights for design thinking practitioners from research," in *Design Thinking Research* (Cham: Springer), 123–145.

Sirkin, D. (2011). "Physicality in distributed design collaboration," in *Design Thinking* (Berlin, Heidelberg: Springer), 165–178.

Sirkin, D., Ju, W., and Cutkosky, M. (2012). "Communicating meaning and role in distributed design collaboration: how crowdsourced users help inform the design of telepresence robotics," in *Design Thinking Research* (Berlin, Heidelberg: Springer), 173–187.

Skulmowski, A., and Xu, K. (2022). Understanding cognitive load in digital and online learning: a new perspective on extraneous cognitive load. *Educ. Psychol. Rev.* 34, 1–26. doi: 10.1007/s10648-021-09624-7

Suleri, S., Sermuga Pandian, V. P., Shishkovets, S., and Jarke, M. (2019). "Eve: a sketch-based software prototyping workbench," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK), 1–6.

Unger, J. M., Bhattarai, A., Maisch, B., Luetz, J. M., and Obuhuma, J. (2021). "Fostering innovation and intercultural exchange during a global pandemic: lessons learned from a virtual design thinking challenge in Nepal," in *COVID-19: Paving the Way for a More Sustainable World* (Cham: Springer), 185–209.

Vallis, C., and Redmond, P. (2021). Introducing design thinking online to large business education courses for twenty-first century learning. *J. Univer. Teach. Learn. Pract.* 18, 213–234. doi: 10.53761/1.18.6.14

Verganti, R., Vendraminelli, L., and Iansiti, M. (2020). Innovation and design in the age of artificial intelligence. *J. Product Innovat. Manag.* 37, 212–227. doi: 10.1111/jpim.12523

Webb, P. (2008). *Extending a Distributed Cognition Framework: The Evolution and Social Organisation of Line Control* (Ph.D. thesis. London University.

Wenzel, M., Gericke, L., Thiele, C., and Meinel, C. (2016). "Globalized design thinking: Bridging the gap between analog and digital for browser-based remote collaboration," in *Design Thinking Research* (Cham: Springer), 15–33.

Yarmand, M., Chen, C., Gasques, D., Murphy, J. D., and Weibel, N. (2021). "Facilitating remote design thinking workshops in healthcare: the case of contouring in radiation oncology," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan), 1–5.

Zhang, J., and Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cogn. Sci.* 18, 87–122. doi: 10.1207/s15516709cog1801_3

Zimmerman, J., and Forlizzi, J. (2008). The role of design artifacts in design theory construction. *Artifact* 2, 41–45. doi: 10.1080/17493460802276893

Zimmerman, J., Forlizzi, J., and Evenson, S. (2007). "Research through design as a method for interaction design research in HCI," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, PA), 493–502.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/frai.2022.919534

OPEN ACCESS

EDITED BY
Sébastien Lallé,
Sorbonne Universités, France

REVIEWED BY
Seiji Yamada,
National Institute of Informatics, Japan
Charles B. Stone,
John Jay College of Criminal Justice, United States

*CORRESPONDENCE
Kyra Göbel
kyra.goebel@fau.de

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 13 April 2022
ACCEPTED 31 October 2022
PUBLISHED 23 November 2022

CITATION
Göbel K, Niessen C, Seufert S and Schmid U (2022) Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. *Front. Artif. Intell.* 5:919534. doi: 10.3389/frai.2022.919534

Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations

Kyra Göbel[1]*, Cornelia Niessen[1], Sebastian Seufert[2] and Ute Schmid[2]

[1]Department of Psychology, Work and Organizational Psychology Unit, Friedrich-Alexander University of Erlangen-Nürnberg, Erlangen, Germany, [2]Information Systems and Applied Computer Science, University of Bamberg, Bamberg, Germany

In the digital age, saving and accumulating large amounts of digital data is a common phenomenon. However, saving does not only consume energy, but may also cause information overload and prevent people from staying focused and working effectively. We present and systematically examine an explanatory AI system (Dare2Del), which supports individuals to delete irrelevant digital objects. To give recommendations for the optimization of related human-computer interactions, we vary different design features (explanations, familiarity, verifiability) within and across three experiments ($N_1 = 61$, $N_2 = 33$, $N_3 = 73$). Moreover, building on the concept of distributed cognition, we check possible cross-connections between external (digital) and internal (human) memory. Specifically, we examine whether deleting external files also contributes to human forgetting of the related mental representations. Multilevel modeling results show the importance of presenting explanations for the acceptance of deleting suggestions in all three experiments, but also point to the need of their verifiability to generate trust in the system. However, we did not find clear evidence that deleting computer files contributes to human forgetting of the related memories. Based on our findings, we provide basic recommendations for the design of AI systems that can help to reduce the burden on people and the digital environment, and suggest directions for future research.

KEYWORDS

distributed cognition, transactive memory, trust, forgetting, explainable AI, human-AI partnership

Introduction

Digital data carriers such as hard drives or cloud spaces have become important memory partners, and cognitive offloading—that is, externally saving information to reduce information processing requirements—can be used to decrease the cognitive demands of a task (Risko and Gilbert, 2016). However, in today's increasingly digitized world of work, individuals save and accumulate more digital objects in their external memory than they actually need in the short and long run. Thus, deleting or archiving irrelevant and outdated data files on a regular basis is important in several respects:

It helps to reduce information overload, limits distractions, enables working in an effective, focused, and goal-oriented manner (Edmunds and Morris, 2000; Hair et al., 2007; Dabbish and Kraut, 2010; Soucek and Moser, 2010; Niessen et al., 2020b), and saves energy (Rong et al., 2016). However, often people do not delete irrelevant files, as deleting tends to be a decision under uncertainty, is effortful, and takes time. Therefore, people might benefit from an AI system designed to support individuals in deleting irrelevant digital objects in external memory (on the computer) on a regular basis. Research has shown that the transparency of system recommendations is important for willingness to use such a system and trust in a system (Pu and Chen, 2007; Wang and Benbasat, 2007; Mercado et al., 2016; Ribeiro et al., 2016; Miller, 2018; Thaler and Schmid, 2021). Thus, the AI system providing explanations plays a central role in the interaction between humans and the AI system.

To investigate whether and how an explanatory interactive AI system helps people to delete irrelevant files from their external memory (i.e., storage device), we developed an assistive AI system (Dare2Del) and conducted three experimental studies focusing on the role of explanations of Dare2Del's recommendations for users' attitudes (information uncertainty, trust), behavior (deleting files) and memory (forgetting irrelevant files). Specifically, building on the concept of distributed cognition (Hutchins, 1995; Zhang and Patel, 2006), which proposes that cognition exists both inside and outside the individual mind, prompting users to actively delete irrelevant digital objects and explaining to them why might also encourage forgetting of the related content in human memory (Sahakyan et al., 2008; Foster and Sahakyan, 2011).

Dare2Del is been developed since 2018 with the main intention to demonstrate how methods of explainable AI can be combined with interactive machine learning to keep humans in the loop in AI supported decision making (Niessen et al., 2020b; Schmid, 2021). As domain, the identification of irrelevant digital objects in the context of work has been selected for several reasons. First, in work contexts, whether a file should be deleted or not is determined by explicit laws and regulations as well as by personal preferences. Therefore, the domain is suitable for AI approaches which combine knowledge-based methods and machine learning (Muggleton et al., 2018). Second, in the context of work erroneous deletion of files might have severe consequences in contrast to private contexts and therefore, the domain is suitable to investigate the effect of explanatory and interactive AI methods on trustworthiness. Third, cloud storing of data comes with high monetary as well as environmental costs and therefore, intelligent tools to identify irrelevant files which can be deleted are of high practical relevance. Over the last years, some products which support the identification of irrelevant files have been developed, mostly in the context of file management systems, some in the context of cloud environments. For instance, Google Photos includes a feature which offers suggestions to delete photos. Suggestions are based on general characteristics such as file size, quality, unsupported format and source. In contrast, Dare2Del can take into account general rules (such as that invoices must be stored for 10 years) as well as individual preferences (such as that for presentation where a pptx exists the pdf can be deleted) which can be given as explicit rules as well as learned from feedback given to suggestions. The tool most similar to Dare2Del is offered by the teaching and learning software Canvas (https://community.canvaslms.com/t5/Canvas-Instructional-Designer/Tool-to-Identify-and-Delete-Unused-Files/ba-p/276260). However, this tool is only designed to delete unused files and empty folders directly from Canvas.

Dare2Del has been explored by five test users who work in the administration of a large company. They used a restricted version of Dare2Del on a file system which has been constructed as a mirror of their own. They used Dare2Del for a month and the general feedback has been positive. However, we are interested in a more controlled evaluation of Dare2Del in an experimental setting. For this reason, a fictitious work context had to be created which does not need specialized knowledge (e.g., accounting). At the same time, the digital objects have to be associated with some relevance such that erroneous deletion would have negative consequences. We decided to use the context of a library system where students' theses have to be archived as a suitable domain which is introduced in detail below.

Our research offers the following contributions to research on human AI collaboration: First, we provide a comprehensive analysis of both behavioral (i.e., accepting the system's suggestions) and cognitive (i.e., trust and credibility building) outcome variables. This allows us to not only identify *if* an assistive system can support users to delete irrelevant files, but also *how* it can help. Thus, our research also offers possible starting points for future improvement and individual or contextual adaptations which can help to increase deleting behavior. This is especially important, as people often do not delete irrelevant or obsolete digital objects in their working and private life. If an explainable AI system can initiate and support behavior change (i.e., lead to increased deleting of files), this might have positive consequences for individuals' stress levels and performance, but also for organizational effectiveness and energy saving. Moreover, we aim to explore underlying mechanisms of action and explain *why* explanations might be beneficial and help to change behavior (i.e., lead to increased deleting of files): We propose that explanations reduce information uncertainty, which in turn leads to more acceptance of the AI system's recommendations and the deletion of the proposed files. An understanding of these mechanisms informs design and interventions to enhance trust, credibility and behavior change in human-AI interaction.

Second, our research adds to the literature on distributed cognition by testing the assumption that an action in external memory (i.e., digital storage devices) has consequences for the

corresponding internal mental representation. Previous research has already shown that external memory is used to store information outside ourselves and that this information is still connected to our memory (Sparrow et al., 2011). However, whether actively deleting information in the external memory can facilitate human forgetting of the connected memory has not been empirically investigated yet. To prove and extend existing research on the theory of distributed cognition, we ask whether deleting a digital object—especially when being convinced about why it should be deleted—also prompts forgetting of the corresponding memory content.

Theoretical background

The role of explanations

An essential prerequisite for cooperative interaction between humans and AI systems is that system decisions are transparent and comprehensible (Muggleton et al., 2018). This requirement is most obvious in the context of machine learning, particularly black-box systems such as deep neural networks. Consequently, explainable AI (XAI; Miller, 2018) has been established as a new area of research, providing methods to make the decisions of machine-trained models more transparent. Several methods for highlighting the relevance of input features have been developed. For instance, visualizing the regions in an input image that had the strongest impact on the classification decision can help to identify overfitting (Lapuschkin et al., 2019). One of the most well-known methods is LIME (Ribeiro et al., 2016)—a model-agnostic method which can be applied not only to image data but also to text. XAI methods providing information about relevance are primarily helpful to model developers, and are often not informative enough for domain experts and do not provide information helpful for end-users (Schmid, 2021). For instance, in medical diagnostics, highlighting might reveal that a model that returns a specific tumor type was right for the wrong reasons because the relevant information used is some textual mark at the image border. For experts, more expressive explanations such as rules or natural language are often more helpful. For instance, the decision between two different severity classes for a tumor might depend on spatial relations such as intrusion into fat tissue or quantifications such as the number of metastases (Bruckert et al., 2020).

Explanations serve to provide reasons for an observed state of affairs (Keil, 2006; Asterhan and Schwarz, 2009; Lombrozo, 2016). Often, causal explanations also serve to justify decisions made, i.e., provide reasons why a decision is "right" (Keil, 2006; Biran and Cotton, 2017). In the field of recommender systems, different types of explanations, in particular feature-based, personalized, and non-personalized explanations, have been identified and empirically investigated in terms of their effectiveness for recommending movies (Tintarev and Masthoff, 2012). In the context of recommender systems, an extensive user study demonstrated that explanations increase willingness to use the system again and that trust in system recommendations reduces cognitive effort (Pu and Chen, 2007).

Distributed cognition

Distributed cognition describes the phenomenon that knowledge exists not only inside the individual, but also in his or her surroundings and within a more complex context—for example, the social, physical, or digital environment (Hutchins, 1995; Zhang and Patel, 2006). These different knowledge domains are interconnected and can influence each other not only in individual, but also in broader collective and cultural contexts (Hoskins, 2016; Sutton, 2016). Therefore, they benefit from being analyzed and treated as a holistic system. Phenomena such as saving-enhanced memory (Storm and Stone, 2015) or the photo-taking impairment effect (Henkel, 2014; Soares and Storm, 2022) show that our digital environments can be used to outsource information and provide cognitive relief (Clark and Chalmers, 1998).

Surprisingly, most basic research on human-computer interaction has not explicitly attempted to investigate conditions and outcomes of these cross-connections and information transfer processes, and the ways that people use external anchors, tools, and storage options to support and relieve their cognitive resources are rather poorly understood (Perry, 2003). We argue that analyzing the connections between internal (i.e., human memory) and external (i.e., computer memory) cognition might not only lead to a better understanding of how the different domains are coordinated and connected; it would also provide an important basis for recommendations on how human-computer interaction processes can be supported. The fact that cognitions are not only distributed, but also connected, makes it possible to determine several starting points for possible interventions. Interventions with respect to dealing with large amounts of data and information overload could start either with the user or with the computer system. For example, a reduction in load could be achieved by deleting files, which externally limits the amount of information, removes potential distractors, organizes the work environment, and therefore contributes to mental relief (Chen et al., 2012). This could further help individuals stop distracting, task-irrelevant thoughts, focus on their actual work tasks, and improve well-being (e.g., Randall et al., 2014; Kluge and Gronau, 2018; Niessen et al., 2020b; Göbel and Niessen, 2021).

In this way, the concept of distributed cognition is important from various perspectives and provides an appropriate framework for comprehensively examining human-computer interactions and developing, designing, and optimizing corresponding assistive systems.

Development of hypotheses and research questions

As causal explanations show that there are relevant and intelligent considerations behind the system's suggestions (e.g., Keil, 2006; Biran and Cotton, 2017), we propose that explanations make it more likely that individuals will delete the proposed files (Hypothesis 1a). Moreover, we aim to replicate the finding that explanations lead to more trust in the system. Trust is defined as the willingness to rely on a technical system in an uncertain environment (Komiak and Benbasat, 2004; Meeßen et al., 2020) and has two components, one of which is more emotional and affective and one of which is more cognitive. *Affective trust* describes the user's feelings while relying on the technical system, whereas *cognitive trust* can be seen as the system's perceived trustworthiness (Komiak and Benbasat, 2006; Meeßen et al., 2020). Both affective and cognitive trust have been shown to have positive effects on intentions to adopt and work with technical agents (Meeßen et al., 2020) as well as on work outcomes and well-being (Müller et al., 2020), and thus should be considered when evaluating such systems. In line with previous research demonstrating that transparency is conducive to the development of trust (e.g., Pu and Chen, 2007; Pieters, 2011; Shin, 2021), we assume that providing explanations increases both affective and cognitive trust ratings (Hypothesis 1b).

Another important factor in this context is credibility. Described as the believability of information and its source (e.g., Fogg et al., 2001), credibility has been identified as one of the strongest predictors of trust in information systems at work (Thielsch et al., 2018). We assume that explanations generally increase the comprehensibility and transparency of the system's decisions. By providing information on *why* the system's suggestions are valid, users can better understand the reliability of the underlying processes. This should lead to increased credibility ratings (Hypothesis 1c).

Furthermore, explanations can reduce information uncertainty (Van den Bos, 2009), here the lack of information about why the system considers a file irrelevant ("why am I getting this particular file suggested"), and thus increase the likelihood of accepting the system's recommendations. As information uncertainty is often experienced negatively (e.g., Wilson et al., 2005) and can lead to ruminative thinking (Kofta and Sedek, 1999; Berenbaum et al., 2008), it might also negatively impact trust and credibility. Therefore, we not only hypothesize that explanations directly reduce information uncertainty (Hypothesis 1d), but also that information uncertainty in the system's proposals mediates the effect of explanations of the system's proposals on its acceptance (Hypothesis 2a), trust (Hypothesis 2b), and credibility (Hypothesis 2c).

It has already been shown that person-situation interactions predict how people deal with too much information in the related field of thought control (Niessen et al., 2020a). Building on these findings, we also assume that there are individual differences in the extent to which explanations support individuals' deletion of irrelevant files, trust in the AI system and finding the suggestions credible. Therefore, we investigated the moderating role of conscientiousness and need for cognition on the relation between explanations and acceptance, trust and credibility. As a personality trait, the need for cognition refers to people's tendency to engage in and enjoy thinking (Cacioppo and Petty, 1982). Individuals with a high need for cognition seek out for information to make sense of stimuli and events. Such individuals enjoy situations in which problem solving and reflection are required (Cacioppo et al., 1996). Therefore, we propose that individuals high in need for cognition have a stronger preference for thinking about the explanations, which helps them to delete irrelevant files (Hypothesis 3a), build trust (Hypothesis 3b) and credibility (Hypothesis 3c) and to reduce information uncertainty (Hypothesis 3d).

Conscientiousness is one of the Big Five personality dimensions (Barrick and Mount, 1991; Costa et al., 1991; Costa and McCrae, 1992). Conscientiousness includes the will to achieve, self-motivation, and efficaciousness, but also a dependability component that is related to orderliness, reliability, and cautiousness. We expect that conscientious individuals read and think about the explanations more deeply, as they are more cautious than less conscientious individuals. Moreover, individuals high in conscientiousness might find the explanations helpful for achieving their work goals, as deleting irrelevant files has positive consequences in terms of reduced information overload, and distractions. Therefore, we propose that conscientiousness moderates the impact of explanations on deletion of irrelevant files (Hypothesis 4a), building trust (Hypothesis 4b), and credibility (Hypothesis 4c), and on reducing information uncertainty (Hypothesis 4d).

We also hypothesize that deletion is not only an action that causes digital objects to be forgotten in external memory, but may also support intentional forgetting of associated memory content (Hypothesis 5; Bjork et al., 1998; Anderson and Hanslmayr, 2014).

Sparrow et al. (2011) showed that individuals were worse at recalling information that had been stored in external memory than information that had not been stored on the computer. This indicates that individuals need to be convinced that they will not need the information designated as irrelevant in the future in order to forget: they need to trust the system. Numerous studies on directed forgetting (Sahakyan et al., 2008; Foster and Sahakyan, 2011) and motivated forgetting (for a review, Anderson and Hanslmayr, 2014) support these assumptions. Here, we explore whether explanations can help users not only delete irrelevant information but also forget it in their memory. When they are informed why a file is irrelevant, individuals can make an informed decision, and if they accept the suggestion to delete, actually forget the file as well (research question).

The present research

We conducted three experiments to test our hypotheses. Participants regularly interacted with the AI system Dare2Del and processed deletion suggestions in all experiments. In the first experiment, we investigated how explanations affect users' attitudes (information uncertainty, trust), behavior (deleting files), and memory (forgetting irrelevant files). In the second experiment, we further enhanced the system's transparency and verifiability by giving users the opportunity to check the correctness of the suggestions. In the third experiment, we additionally varied memory processing depth of the to-be-deleted material to further elaborate and systematize effects on memory.

To estimate the required sample size, we conducted multilevel power analyses for cross-level interaction effects (Multilevel Power Tool; Mathieu et al., 2012). We elected to use an anticipated effect size in the small-to-medium range (0.25, $p = 0.05$, Power 95%) and followed parameter recommendations from Mathieu et al. (2012) and Arend and Schäfer (2019) and in order to conservatively estimate our sample sizes. Moreover, Experiment 1 (https://aspredicted.org/3YN_FYC) and Experiment 3 (https://aspredicted.org/MQT_TYG) were preregistered on aspredicted.org. All data are publicly available on OSF (https://osf.io/dk6en/).

Experiment 1

Method

Participants

The study was conducted with 61 undergraduates (majoring in psychology; 49 female, 12 male) from a German university. Mean age was 20.30 years ($SD = 3.00$, range 18–32). Participation was voluntary and participants received course credit as compensation.

Materials and procedure

The experiment was programmed with SoSciSurvey software (with additional php elements), conducted online, and lasted about an hour. During the experiment, participants were in contact with the experimenter via video chat. Before starting the experiment, demographic information (age, gender, and occupation), need for cognition, and conscientiousness were assessed with a questionnaire.

At the beginning of the experiment, participants were instructed that they would be testing a library system (see Supplementary material) at the university which digitally saves and manages dissertations, diploma, bachelor's, and master's theses (cover story). First, the participants' main task was to archive students' theses. Specifically, they had to process 36 emails from students who had sent their theses along with Supplementary information (short cover letter including author name, name of thesis, type of thesis, publication year). Participants then entered the important meta-information (title, author name, type of thesis, publication year) from all 36 emails into the digital library system and saved the email attachments automatically by pressing the respective button. After each email, they received brief feedback from the system that the entry had been saved. The emails were presented in random order. To ensure that archiving the theses always involved a comparable workload, all titles had a similar structure and consisted of two technical terms (e.g., "neuroticism and burnout").

Second, participants were instructed to interact with an assistive system (Dare2Del) that helps to keep the digital library system tidy, without outdated or duplicate theses. The assistive system Dare2Del was described as an automatic software that detects irrelevant, identical and damaged files and suggests them for deletion. Participants were explicitly advised that the decision on whether to accept or reject the suggestion was completely up to them. Nevertheless, they were also encouraged to keep the archival system organized by using Dare2Del. While processing the emails, the assistive system popped up 12 times. Each time, a file was presented and suggested for deletion (see Supplementary material).

We systematically varied the explanation for why the file should be deleted: Six files were suggested without explanation, and for six files the assistive system provided a short explanation (three different explanations: thesis file is identical to another and was obviously saved twice; thesis file is outdated, and a newer version exists; thesis file was submitted at another university and therefore should not be in the system). Also, we systematically varied the familiarity of the files. Six of the to-be-deleted files were thesis files that the participants had previously saved into the archival system—that is, they had already entered the thesis titles into the system and saved the respective information. Six files, on the other hand, were completely new files that had not been presented before (unfamiliar files).

After processing all emails from students and suggestions for deletion from the assistive system, participants completed a recognition test. The 12 thesis file names the assistive system had suggested during the experiment (e.g., "narcissism and loneliness") and 12 distractors (thesis file names with slightly modified titles; e.g., "egoism and loneliness") were presented to the participants in a random order. Participants were asked to indicate whether they had processed the this exact title before or not. In this way, we assessed whether participants could correctly identify the original files they had dealt with before. At the end of the experiment, participants had the opportunity to make general comments and were then fully debriefed.

Research design

The experimental design included an explanation condition (suggestion with explanation, coded as 1, and without explanation, coded as 0; within-person) and need for cognition and conscientiousness (between-person, see Figure 1). Need for

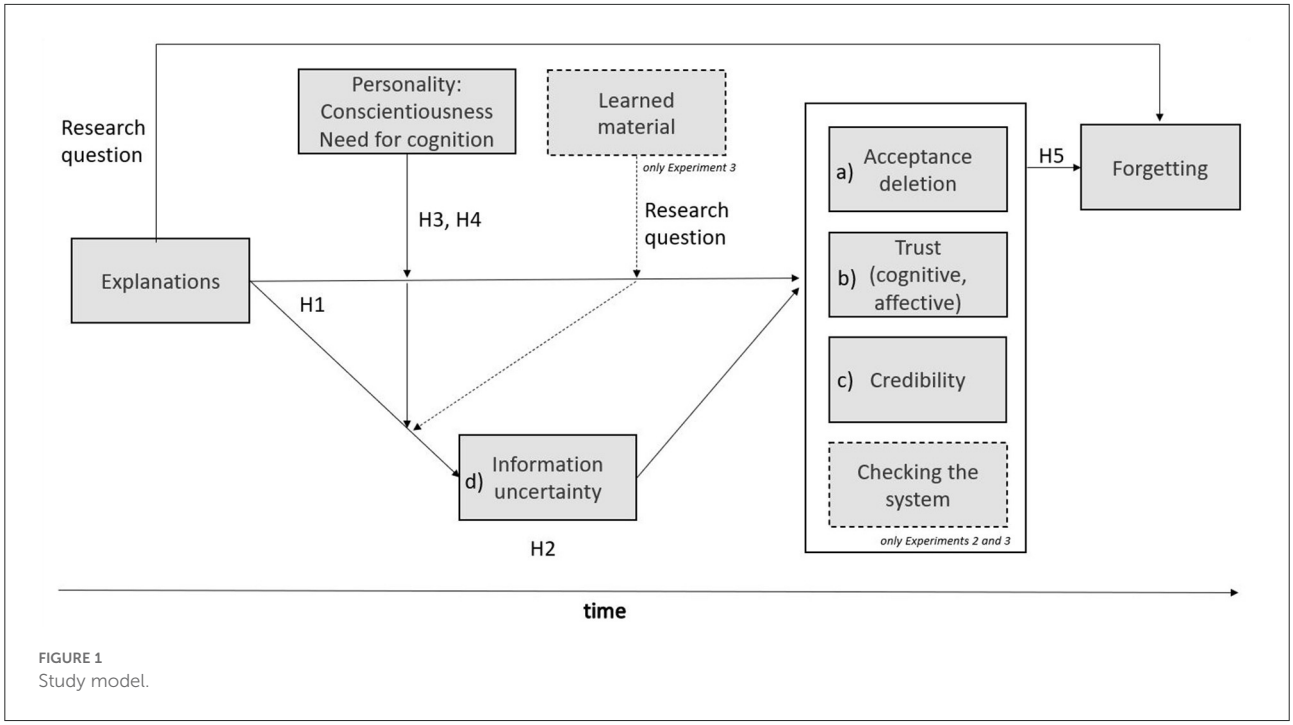

FIGURE 1
Study model.

cognition was assessed with 33 items (e.g., "I really enjoy the task of finding new solutions for problems"; Bless et al., 1994; Cronbach's Alpha = 0.92), and conscientiousness with the NEO-PI-R (60 items; Ostendorf and Angleitner, 2004; e.g., "I work goal-oriented and effectively", Cronbach's Alpha = 0.85).

Dependent variables

We assessed five dependent variables. Firstly, we recorded whether the assistive system's suggestion to delete the file was accepted or not (no = 0; yes = 1). Secondly, after participants accepted or refused the suggestion, we measured trust with two components, namely cognitive trust ("I feel comfortable relying on the assistive system") and affective trust ("I trust the assistive system completely"). The third dependent variable we measured was credibility ("The information given by the assistive system was credible") and the fourth was information uncertainty ("I feel uncertain as to why the file was suggested for deletion, because I do not have enough information"). Trust, credibility, and information uncertainty were answered on a 5-point Likert scale ranging from 1 = *do not agree at all* to 5 = *fully agree*. Finally, we assessed the hit rate of the thesis names in the recognition test (no = 0; yes = 1).

Control variables

As control variable, we assessed the familiarity of the theses' titles (familiar, coded as 1, files processed by the participants; unfamiliar, coded as 0, new files not presented to participants). To consider possible effects of practice with the task, we further included a time variable in the model representing the position of Dare2Del's suggestion to delete a file (0–11). This variable makes it possible to detect systematic changes over time.

Results

Statistical analyses

Multilevel modeling and logistic multilevel modeling were used to conduct the within-person comparison of experimental conditions. Multilevel modeling presents a valuable alternative approach to traditional repeated measures analysis of variance (RM-ANOVA), as it is more robust to violations of assumptions, can handle missing data, and allows for testing more complex hierarchical structures (Cohen et al., 2003). We used R software and the packages *lme4* (Bates et al., 2015) and *mediation* (Tingley et al., 2014) to conduct our analyses. All models were two-level models, with suggestions by Dare2Del with and without explanations, familiarity of the theses' titles, deletion decisions and recognition of files in the final recognition test nested within individuals at Level 2.

The continuous Level 1 (within) predictor variable information uncertainty was centered around the person mean (Nezlek, 2012), and the continuous Level 2 (between) predictor variables (need for cognition, conscientiousness) were centered around the grand mean. Dummy-coded predictor variables were entered uncentered, as were all outcome variables for the respective models.

We applied the two-step approach to causal mediation analysis documented by Imai et al. (2010) and Tingley et al. (2014): In the first step, the mediator variable is predicted by the

predictor variable, and in the second step, the outcome variable is predicted by the predictor and mediator variables. The final mediation analysis is then run using quasi-Bayesian Monte-Carlo simulations (we used 10,000 simulations each), which is superior to previous mediation approaches as it overcomes problems such as dependence on specific statistical models or restrictive assumptions (cf. Imai et al., 2010; Pearl, 2014; Tingley et al., 2014).

$N_{Level2}$was 61, the number of the participants. Due to technical problems, we had to exclude two Level 1 datapoints, leading to an $N_{Level1}$of 730 (61 participants × 12 deletion proposals-−2 excluded datapoints). Overall, participants accepted about one third (36%) of the system's deletion proposals. In the final recall test, 48% of the files were identified correctly. An overview of all Level 1 and Level 2 variable correlations is provided in Tables 1, 2, respectively.

Hypothesis testing

First, we tested the hypothesis that explanations lead to higher acceptance of the assistive system's suggestions, to more cognitive and affective trust, more credibility, and less information uncertainty (Hypotheses 1a–d). To do so, we calculated (logistic) multilevel regression analyses. In line with our expectations, the presence of explanations led to higher acceptance of the deletion suggestions ($\gamma = 3.96$, $SE = 0.31$, $z = 12.63$, $p < 0.001$). However, explanations did not increase trust (cognitive trust: $\gamma = -0.02$, $SE = 0.06$, $t = -0.37$, $p = 0.709$; affective trust: $\gamma = -0.02$, $SE = 0.07$, $t = -0.24$, $p = 0.814$) or the credibility of the system ($\gamma = 0.02$, $SE = 0.08$, $t = 0.20$, $p = 0.840$). Contrary to our expectations, explanations for the suggestions increased rather than decreased information uncertainty ($\gamma = 0.21$, $SE = 0.09$, $t = 2.30$, $p = 0.021$). However, it should be noted that due to simultaneous testing of up to five dependant variables, the p-value needs to be adjusted down to 0.01 (0.05/5; Haynes, 2013).

The familiarity of the files, which we added as a control variable to our analyses to examine possible effects of different levels of cognitive processing, led to more cognitive and affective trust, more credibility and less uncertainty (see Table 3). However, it did not affect acceptance of the suggestions. The results of the time variable revealed a decrease in cognitive and affective trust and an increase in information uncertainty over time (see also Table 3). These findings are somewhat unexpected and need to be further examined and discussed.

Second, we tested the mediating role of information uncertainty with regard to acceptance of the system's suggestions, cognitive trust, affective trust, and credibility. Information uncertainty did not mediate the effect of explanations on acceptance of the deletion suggestions (indirect effect = −0.00, 95% CI [−0.01; 0.00], $p = 0.310$). However, we found indirect effects for cognitive trust (indirect effect = −0.07, 95% CI [−0.13; −0.01], $p = 0.020$), affective trust (indirect effect = −0.09, 95% CI [−0.16; −0.01], $p = 0.020$), and credibility (indirect effect = −0.12, 95% CI [−0.22; −0.02], $p = 0.020$), but as with the results for Hypothesis 1, the direction of effects was unexpected: Explanations created *more* information uncertainty, which resulted in *less* cognitive trust, *less* affective trust, and *less* credibility. Thus, Hypotheses 2a–c were not confirmed, although they highlighted the mediating role of information uncertainty.

In the next step, we tested whether need for cognition (Hypotheses 3a–d) and conscientiousness (Hypotheses 4a–d) moderated the effect of explanations on acceptance, cognitive trust, affective trust, credibility, and information uncertainty. To do so, we calculated cross-level interactions. However, none of them turned out to be significant for either need for cognition (all $zs/ts < |1.14|$, all $ps > 0.252$) or conscientiousness (all $zs/ts < |1.28|$, all $ps > 0.202$) as a moderator. Therefore, we had to completely reject Hypothesis 3 and Hypothesis 4. We further investigated whether deleting a file led to subsequent forgetting. Confirming Hypothesis 5, deleting a file was associated with a lower recognition probability ($\gamma = -1.69$, $SE = 0.19$, $z = -8.69$, $p < 0.001$).

To explore possible effects of explanations on the subsequent accessibility of the corresponding memory traces (research question), we calculated additional multilevel regression analyses. The results revealed that the presence of explanations for a suggestion to delete a thesis file was associated with a lower probability of recognizing its title in the recognition test ($\gamma = -2.17$, $SE = 0.22$, $z = -9.68$, $p < 0.001$), thus indicating difficulties in retrieval (which corresponds to forgetting).

Additional analyses

As explanations were positively associated with deleting a file (Hypothesis 1a), we further conducted a two-step causal mediation analysis to test whether the act of deletion mediates the effect of explanations on subsequent forgetting. The indirect effect was not significant (indirect effect = −0.04, 95% CI [−0.09; 0.01], $p = 0.110$), but the direct effect from explanations on the recognition rate was again confirmed (direct effect = −0.46, 95% CI [−0.54; −0.37], $p < 0.001$). Unexpectedly, time and familiarity of files did not affect final recall rates.

In sum, as expected, explanations led to higher acceptance of the deletion suggestions and to more forgetting of the files. Contrary to our hypotheses, explanations did not increase trust or credibility of the system, but increased information uncertainty, which led to less trust. Moreover, over the course of the experiment, trust actually decreased and information uncertainty increased.

Experiment 2

To further explore the surprising effects of explanations on trust, credibility, and information uncertainty, we conducted a second experiment with two major changes. First, we assumed

TABLE 1 Experiment 1: Means, standard deviations, and correlations of Level 1 variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. Familiarity[a]	0.50	0.50								
2. Explanation[a]	0.50	0.50	0.00							
3. Time[b]	5.49	3.46	−0.05	0.05						
4. Deleted[a]	0.36	0.48	0.02	0.59***	−0.01					
5. Cognitive trust	2.31	1.03	0.12**	−0.02	−0.20***	0.16***				
6. Affective trust	2.38	1.09	0.13***	−0.02	−0.18***	0.17***	0.80***			
7. Credibility	2.82	1.22	0.16***	−0.01	−0.38***	0.09*	0.61***	0.64***		
8. Uncertainty	3.68	1.39	−0.28***	0.09*	0.26***	−0.09*	−0.56***	−0.61***	−0.67***	
9. Recognition[a]	0.48	0.50	0.05	−0.50***	0.01	−0.33***	0.04	0.02	−0.01	−0.02

$N_{Level1} = 730$.
[a]Dichotomous variable: "no" coded as 0, "yes" coded as 1.
[b]Position of deleting proposals (0–11).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 2 Experiment 1: Means, standard deviations, and correlations of Level 2 and aggregated Level 1 variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Age	20.49	3.00									
2. Gender[a]	0.20	0.40	−0.07								
3. Conscientiousness	3.72	0.34	0.03	−0.00							
4. Need for cognition	3.39	0.52	0.13	0.21	0.20						
5. Deleted[b,c]	0.36	0.18	−0.06	−0.06	0.11	−0.21					
6. Cognitive trust[c]	2.30	0.67	−0.01	−0.01	0.28*	−0.22	0.65***				
7. Affective trust[c]	2.37	0.65	−0.04	−0.02	0.22	−0.23	0.74***	0.91***			
8. Credibility[c]	2.82	0.51	−0.01	−0.04	0.12	−0.24	0.51***	0.53***	0.61***		
9. Uncertainty[c]	3.69	0.53	0.22	−0.05	−0.10	0.24	−0.69***	−0.67***	−0.71***	−0.54***	
10. Recognition[b,c]	0.48	0.16	−0.04	−0.06	−0.08	0.19	0.10	0.09	0.10	−0.03	0.09

$N_{Level2} = 61$.
[a]Female coded as 0, male coded as 1.
[b]Dichotomous variable: "no" coded as 0, "yes" coded as 1.
[c]Level 1 variable aggregated on the person-level.
*$p < 0.05$, ***$p < 0.001$.

TABLE 3 Experiment 1: Effects of explanations on acceptance of deleting proposal, cognitive trust, affective trust, credibility, and uncertainty.

	Acceptance of deleting proposal[a]			Cognitive trust[b]			Affective trust[b]			Credibility[b]			Uncertainty[b]		
Predictor	***Est.***	***SE***	***z***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***
Constant	−3.02	0.36	−8.37***	2.53	0.11	23.87***	2.55	0.11	23.50***	3.37	0.11	31.99***	3.42	0.12	28.90***
Time	−0.05	0.03	−1.63	−0.06	0.01	−7.00***	−0.06	0.01	−5.88***	−0.13	0.01	−11.69***	0.10	0.01	7.31***
Familiarity	0.20	0.22	0.90	0.22	0.06	3.73***	0.27	0.07	4.08***	0.34	0.08	4.37***	−0.74	0.09	−8.12***
Explanation	3.96	0.31	12.63***	−0.02	0.06	−0.37	−0.02	0.07	−0.24	0.02	0.08	0.20	0.21	0.09	2.30*

$N_{Level1} = 730$, $N_{Level2} = 61$.
[a]Logistic multilevel regression analysis.
[b]Continuous multilevel regression analysis.
***$p < 0.001$, *$p < 0.05$.

that the negative effects of explanations on trust, credibility, and information security in Experiment 1 were because participants were not able to check the explanations and suggestions in the file system. As a result, they simply accepted the suggestions blindly, but did not trust them, did not find the system credible, and felt more uncertain. In Experiment 2, we provided the

possibility to check the explanations and suggestions by looking at the files in the folder (see Supplementary material). Therefore, we were able to assess participants' trust in a more objective manner (with less checking of explanations indicating more trust). Second, we modified the kind of explanations, so that acceptance of the suggestion would imply definite and final removal (in contrast to the deletion of duplicates, where one of the original files continues to exist in the file system). Therefore, we partly changed the explanations' content (e.g., thesis was submitted at a foreign university that was not part of the literature network system).

Method

Participants

Participants were 33 undergraduates (participating in return for course credit) at a German university (27 female, five male, one non-binary). Mean age was 22.55 years ($SD = 4.98$, range 18–40).

Materials and procedure

The materials and procedure for the experiment were similar to Experiment 1, with two exceptions. First, participants had access to the underlying file system. They could scroll through the file list, which consisted of 50 alphabetically sorted thesis data files, and could check whether the explanations provided by the assistive system were appropriate (e.g., file was duplicate in the system). The explanations were always correct and consistent with the file system. Second, we varied the kind of explanations. One explanation stated that the file had accidently been saved twice, and that because of this, one copy had to be removed. The other explanation stated that the file was erroneously in the file system as it was submitted at a foreign university that was not part of the literature network system, and therefore suggested final removal.

Research design

The design was the same as for Experiment 1.

Dependent variables

In addition to the dependent variables in Experiment 1, we were able to assess an additional measure of trust, namely, whether participants checked the validity of the suggestions. We measured whether participants had opened the underlying file system (no = 0; yes = 1) and how much time the participants spent scrolling through and checking it (in milliseconds).

Control variables

As in Experiment 1, familiarity and time were included as control variables.

Results

Statistical analyses

We followed the same analytic strategy as Experiment 1. $N_{Level2}$was 33, the number of the participants. $N_{Level1}$was 396 (33 participants × 12 deletion proposals). An overview of the correlations of all Level 1 and Level 2 variables is provided in Tables 4, 5, respectively. Overall, participants checked the file system in 70% of all cases, and accepted about two thirds (70%) of the system's deletion proposals. In the final recall test, 43% of the files were identified correctly.

Hypothesis testing

To test Hypotheses 1a–d, we again analyzed the effect of explanations on the acceptance of deletion proposals, cognitive trust, affective trust, checking the file system, credibility, and information uncertainty. Results showed that when explanations were given, participants were more likely to delete the suggested file ($\gamma = 0.53$, $SE = 0.25$, $z = 2.16$, $p < 0.05$), considered the system more credible ($\gamma = 0.29$, $SE = 0.10$, $t = 2.89$, $p < 0.01$), and reported less uncertainty ($\gamma = -0.40$, $SE = 0.12$, $t = -3.33$, $p < 0.001$). They did not trust the system more either cognitively or affectively and there were no effects on frequency of or time spent checking the file system. However, further analyses indicated an increase in affective trust as well as less and shorter periods of checking the file system over time (see Tables 6A,B). Thus, Hypothesis 1 could only be partly confirmed. Familiarity of files did not exhibit any effects.

In the next step, we again tested for the possible mediating role of information uncertainty (Hypotheses 2a–c). Significant mediation processes could be identified for all dependent variables: Explanations generally reduced information uncertainty, and reduced uncertainty in turn led to increased acceptance of deletion proposals (indirect effect = 0.05, 95% CI [0.02; 0.08], $p < 0.001$), more cognitive trust (indirect effect = 0.16, 95% CI [0.06; 0.25], $p < 0.001$), affective trust (indirect effect = 0.16, 95% CI [0.07; 0.26], $p < 0.001$), and credibility (indirect effect = 0.18, 95% CI [0.07; 0.30], $p < 0.001$). No indirect effect of information uncertainty was found for either opening the file system (indirect effect = 0.01, 95% CI [−0.00; 0.02], $p = 0.150$) or time spent checking the file system (indirect effect = 0.05, 95% CI [−0.02; 0.15], $p = 0.180$).

Concerning possible moderating effects of need for cognition (Hypotheses 3a–d) on acceptance of the proposals, cognitive and affective trust, checking the file system, credibility and information uncertainty, we found no significant interactions between presence of explanations and need for cognition on credibility, all $zs/ts < |1.47|$, all $ps > 0.142$.

For conscientiousness (Hypotheses 4a–d), a significant interaction effect with presence of explanations on information uncertainty was found ($\gamma = 0.87$, $SE = 0.35$, $t = 2.52$, $p = 0.012$; see Figure 2): People with lower (−1 SD) conscientiousness reported less information uncertainty when an explanation was

TABLE 4 Experiment 2: Means, standard deviations, and correlations of Level 1 variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Familiarity[a]	0.50	0.50										
2. Explanation[a]	0.50	0.50	0.00									
3. Time[b]	5.50	3.46	−0.05	0.05								
4. Checked[a]	0.70	0.46	0.03	−0.02	−0.07							
5. Checked (time)[c]	6.67	4.40	0.03	−0.01	−0.10*	1.00***						
6. Deleted[a]	0.70	0.46	−0.02	0.09	−0.04	0.17**	0.16**					
7. Cognitive trust	2.94	1.32	0.01	0.06	0.07	0.02	−0.00	0.52***				
8. Affective trust	3.04	1.32	0.00	0.07	0.08	0.06	0.04	0.49***	0.87***			
9. Credibility	3.66	1.31	0.03	0.11*	−0.03	0.22***	0.20***	0.66***	0.67***	0.72***		
10. Uncertainty	2.52	1.53	−0.01	−0.13**	−0.05	−0.28***	−0.26***	−0.60***	−0.52***	−0.52***	−0.61***	
11. Recognition[a]	0.43	0.50	0.05	0.11*	0.36***	0.04	0.03	−0.05	−0.04	−0.05	−0.06	0.01

$N_{Level1} = 396$.
[a]Dichotomous variable: "no" coded as 0, "yes" coded as 1.
[b]Position of deleting proposals (0–11).
[c]Logarithmized, in milliseconds.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 5 Experiment 2: Means, standard deviations, and correlations of Level 2 and aggregated Level 1 variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Age	22.55	4.98											
2. Gender[a]	0.16	0.37	−0.09										
3. Conscientiousness	3.62	0.35	−0.14	0.20									
4. Need for cognition	3.34	0.45	−0.15	0.01	0.47**								
5. Checked[b,c]	0.70	0.40	−0.07	0.19	0.30	0.32							
6. Checked (time)[c,d]	6.67	3.83	−0.07	0.18	0.29	0.31	1.00***						
7. Deleted[b,c]	0.70	0.24	−0.02	−0.01	−0.03	−0.05	0.36*	0.36*					
8. Cognitive trust[c]	2.94	1.00	−0.14	0.16	−0.00	−0.10	0.10	0.08	0.51**				
9. Affective trust[c]	3.04	1.04	−0.17	0.28	−0.05	−0.14	0.11	0.09	0.46**	0.94***			
10. Credibility[c]	3.66	0.88	−0.14	0.29	0.08	−0.03	0.36*	0.35*	0.75***	0.67***	0.74***		
11. Uncertainty[c]	2.52	1.02	−0.06	−0.13	−0.03	−0.25	−0.44*	−0.43*	−0.69***	−0.52**	−0.50**	−0.68***	
12. Recognition[b,c]	0.43	0.18	−0.32	−0.15	0.26	0.24	0.21	0.22	−0.06	−0.20	−0.32	−0.23	0.22

$N_{Level2} = 33$.
[a]Female coded as 0, male coded as 1.
[b]Dichotomous variable: "no" coded as 0, "yes" coded as 1.
[c]Level 1 variable aggregated on the person-level.
[d]Logarithmized, in milliseconds.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

provided (simple slope = −0.70, $t = -4.15$, $p < 0.001$). For people with higher (+1 *SD*) conscientiousness, no difference was found (simple slope = −0.09, $t = -0.54$, $p = 0.591$). For all other dependent variables, no effects were detected, all $zs/ts < |1.88|$, all $ps > 0.061$. Therefore, Hypothesis 4 could only be confirmed for information uncertainty.

Lastly, we tested whether explanations and the deletion of files led to more subsequent difficulties in recognizing the thesis titles (H5). Contrary to the results of Experiment 1, we found that explanations were associated with a higher likelihood of subsequent recognition ($\gamma = 0.46$, $SE = 0.23$, $z = 2.00$, $p = 0.046$), and deleting a file was not related to recognition at all ($\gamma = -0.24$, $SE = 0.27$, $z = -0.92$, $p = 0.358$). Therefore, Hypothesis 5 was not supported.

Additional analyses

We tested for differences between the two kinds of explanations. The results revealed that participants accepted more suggestions to delete duplicates ($\gamma = 2.43$, $SE = 0.46$, $z = 5.26$, $p < 0.001$) and fewer suggestions to delete files that were erroneously in the system ($\gamma = -0.64$, $SE = 0.31$, $z = -2.08$, p

TABLE 6A Experiment 2: Effects of explanations on acceptance of deleting proposal, cognitive trust, affective trust, credibility, and uncertainty.

	Acceptance of deleting proposal[a]			Cognitive trust[b]			Affective trust[b]			Credibility[b]			Uncertainty[b]		
Predictor	***Est.***	***SE***	***z***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***
Constant	1.16	0.38	3.07**	2.71	0.20	13.58***	2.79	0.20	13.78***	3.54	0.19	18.87***	2.86	0.22	13.05***
Time	−0.04	0.04	−1.02	0.02	0.01	1.87	0.03	0.01	2.27*	−0.01	0.01	−0.87	−0.02	0.02	−1.24
Familiarity	−0.11	0.25	−0.44	0.04	0.09	0.42	0.01	0.09	0.17	0.08	0.10	0.81	−0.05	0.12	−0.40
Explanation	0.53	0.25	2.16*	0.14	0.09	1.57	0.17	0.09	1.92	0.29	0.10	2.89**	−0.40	0.12	−3.33***

$N_{Level1} = 396$, $N_{Level2} = 33$.
[a]Logistic multilevel regression analysis.
[b]Continuous multilevel regression analysis.
***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

TABLE 6B Experiment 2: Effects of explanations on checking the system.

	Checking the file system[a]			Time spent checking the file system[b,c]		
Predictor	***Est.***	***SE***	***z***	***Est.***	***SE***	***t***
Constant	4.27	2.07	2.06*	7.28	0.71	10.24***
Time	−0.16	0.06	−2.44*	−0.13	0.03	−3.79***
Familiarity	0.38	0.43	0.90	0.23	0.23	0.97
Explanation	−0.22	0.42	−0.51	−0.04	0.23	−0.17

$N_{Level1} = 396$, $N_{Level2} = 33$.
[a]Logistic multilevel regression analysis.
[b]Continuous multilevel regression analysis.
[c]Logarithmic (originally in milliseconds).
***$p < 0.001$, *$p < 0.05$.

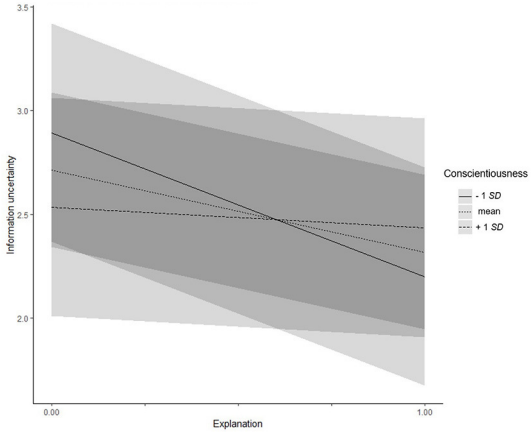

FIGURE 2
Experiment 2: Interaction effect between conscientiousness and explanations on information uncertainty. Low and high levels of information uncertainty represent one standard deviation below and above the mean, respectively.

< 0.05) compared to files with no explanations. There was no effect of explanations on checking the file system. However, files with the duplicate explanation were more likely to be recognized in the recall test than files with no explanation ($\gamma = 0.68$, $SE = 0.31$, $z = 2.23$, $p < 0.05$), whereas there was no difference between files identified as erroneously in the system and the no explanation condition.

In sum, in this experiment, the possibility of checking why the system suggested a thesis for deletion led not only to more suggestions being accepted, but also to more trust over time, credibility and information uncertainty. Explanations reduced information uncertainty, which was in turn related to more trust and credibility. However, in contrast to Experiment 1, explanations also improved recall of the titles suggested for deletion, and did not promote forgetting.

Experiment 3

The aim of Experiment 3 was to replicate the results of Experiment 2 and investigate whether explanations promote forgetting of well-known information (i.e., thesis titles). Based on research on intentional forgetting, we assumed that an explanation for why a file can be deleted should indicate to participants that the memory content connected to this file can be intentionally forgotten ("I don't need it anymore, so I can forget it"). In the previous experiments, however, we did not control for whether our participants had actually remembered the thesis titles they entered into the database. Therefore, in this experiment, one group of participants had to learn and remember the thesis titles.

Method

Participants

Experiment 3 was conducted with 73 undergraduates (55 female, 18 male) at a German university. Mean age was 23.14 years ($SD = 3.87$, range 18–36). Participation was voluntary and

TABLE 7 Experiment 3: Means, standard deviations, and correlations of Level 1 variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Familiarity[a]	0.50	0.50										
2. Explanation[a]	0.50	0.50	0.00									
3. Time[b]	5.50	3.45	−0.05	0.05								
4. Checked[a]	0.71	0.45	0.04	−0.07*	−0.05							
5. Checked (time)[c]	6.59	4.22	0.04	−0.07*	−0.09**	0.99***						
6. Deleted[a]	0.68	0.47	−0.01	0.18***	0.15***	0.14***	0.13***					
7. Cognitive trust	2.75	1.21	−0.03	0.11**	0.21***	−0.04	−0.06	0.52***				
8. Affective trust	2.84	1.22	0.02	0.12***	0.20***	0.02	−0.01	0.52***	0.84***			
9. Credibility	3.60	1.12	0.01	0.20***	0.16***	0.13***	0.11**	0.49***	0.52***	0.55***		
10. Uncertainty	2.87	1.44	−0.06	−0.21***	−0.22***	−0.07*	−0.05	−0.55***	−0.54***	−0.56***	−0.44***	
11. Recognition[a]	0.49	0.50	0.06	0.09**	0.40***	0.06	0.05	0.06	0.03	0.04	0.07	−0.07*

$N_{Level1} = 876$.
[a]Dichotomous variable: "no" coded as 0, "yes" coded as 1.
[b]Position of deleting proposals (0-11).
[c]Logarithmized, in milliseconds.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

participants received either course credit or a financial reward (€ 15) as compensation.

Materials and procedure

The experimental task was similar to that in Experiment 2 but consisted of two parts. In the learning phase, participants ($n = 32$) in the learning condition were instructed to learn the six familiar file names. To ensure that the file names had been learned sufficiently, a recognition test with the file names as well as six distractor names was conducted. Only if there were no recognition errors did the main part of the experiment—archiving theses—start; otherwise, participants had to relearn the thesis titles and were then given a second recognition test. In the control condition, participants ($n = 41$) did not learn the file names before starting the main part of the experiment. The materials and procedure for the main part of the experiment were the same as in Experiment 2. However, for reasons of consistency, we used two explanations from Experiment 1, neither of which implied final and definitive removal in case of deletion (thesis file is identical to another one and was clearly saved twice; thesis file is outdated and a newer version exists).

Research design

The experimental design included an explanation condition (with explanation, coded as 1; without explanation, coded as 0; within-person), a learning condition (learning coded as 1, and no learning coded as 0; between-person) and need for cognition and conscientiousness (between-person, see Figure 1).

Dependent variables

Dependent variables and measures were the same as in Experiment 2.

Control variables

As in Experiment 2, familiarity and time were included as control variables.

Results

Statistical analyses

We followed the same analytic strategy as Experiments 1 and 2. N_{Level2} was 73, the number of the participants. N_{Level1} was 876 (73 participants × 12 deletion proposals). An overview of the correlations among all Level 1 and Level 2 variables is provided in Tables 7, 8, respectively. Overall, participants checked the file system in 71% of all cases, and accepted about two-thirds (68%) of the system's deletion suggestions. In the recognition test, 49% of the files were identified correctly. In the learning group, 76% of the files were checked, 70% were deleted, and the overall recognition rate was 54%. In the control group, participants checked 68% of the files, deleted 66%, and identified 44% of the files correctly on the final recognition test.

Hypothesis testing

Hypotheses 1a–d proposed that explanations would lead to higher acceptance of the deletion proposals, more trust (cognitive trust, affective trust, and less verification of the suggestions), greater credibility, and less information uncertainty. The results revealed that when explanations were given, participants were more likely to delete a file ($\gamma = 1.13$, $SE = 0.18$, $z = 6.23$, $p < 0.001$), trusted the system more cognitively ($\gamma = 0.23$, $SE = 0.06$, $t = 4.10$, $p < 0.001$) as well as affectively ($\gamma = 0.27$, $SE = 0.06$, $t = 4.88$, $p < 0.001$), checked the file system less frequently ($\gamma = -0.79$, $SE = 0.24$, $z = -3.26$, $p = 0.001$), spent less time checking the system ($\gamma = -0.54$, $SE = 0.19$, $t = -2.92$, $p = 0.004$), considered the system more credible ($\gamma = 0.43$, $SE = 0.06$, $t = 7.58$, $p < 0.001$), and reported less information uncertainty ($\gamma = -0.58$, $SE = 0.07$, $t = -7.50$, $p < 0.001$). Thus, Hypothesis 1 was supported. The results did not differ between participants who had learned the file names prior to the experiment and those who had not. Also, the familiarity of file names, that is, the names of files participants had

TABLE 8 Experiment 3: Means, standard deviations, and correlations of Level 2 and aggregated Level 1 variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Age	23.14	3.87											
2. Gender[a]	0.25	0.43	0.30**										
3. Conscientiousness	3.71	0.39	0.10	0.03									
4. Need for cognition	3.42	0.40	0.09	0.31**	0.11								
5. Checked[b,c]	0.71	0.36	0.14	0.20	−0.24*	−0.09							
6. Checked (time)[c,d]	6.59	3.29	0.15	0.21	−0.24*	−0.09	1.00***						
7. Deleted[b,c]	0.68	0.27	−0.03	−0.01	−0.13	−0.19	0.31**	0.32**					
8. Cognitive trust[c]	2.75	0.87	0.08	−0.16	−0.12	−0.03	−0.00	−0.01	0.54***				
9. Affective trust[c]	2.84	0.89	0.11	−0.12	−0.15	−0.01	0.08	0.07	0.57***	0.91***			
10. Credibility[c]	3.60	0.74	0.10	0.10	0.04	−0.01	0.32**	0.32**	0.55***	0.45***	0.48***		
11. Uncertainty[c]	2.87	0.84	0.05	0.12	0.02	0.09	−0.20	−0.19	−0.63***	−0.49***	−0.49***	−0.39**	
12. Recognition[b,c]	0.49	0.19	−0.12	−0.07	−0.20	−0.18	0.29*	0.31**	0.03	−0.15	−0.07	−0.09	−0.03

$N_{Level2} = 73$.
[a]Female coded as 0, male coded as 1.
[b]Dichotomous variable: "no" coded as 0, "yes" coded as 1.
[c]Level 1 variable aggregated on the person-level.
[d] Logarithmized, in milliseconds.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 9A Experiment 3: Effects of explanations on acceptance of deleting proposal, cognitive trust, affective trust, credibility, and uncertainty.

	Acceptance of deleting proposal[a]			Cognitive trust[b]			Affective trust[b]			Credibility[b]			Uncertainty[b]		
Predictor	***Est.***	***SE***	***z***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***	***Est.***	***SE***	***t***
Constant	−0.15	0.36	−0.42	2.24	0.15	14.98***	2.30	0.15	15.06***	3.09	0.13	23.56***	3.83	0.15	24.79***
Learning	0.26	0.46	0.55	0.04	0.21	0.85	−0.07	0.21	−0.33	0.04	0.18	0.15	−0.21	0.20	−1.07
Time	0.13	0.03	4.99***	0.07	0.01	8.79***	0.07	0.01	8.72***	0.05	0.01	5.84***	−0.09	0.01	−7.93***
Familiarity	−0.03	0.18	−0.17	−0.04	0.06	−0.76	0.08	0.06	1.48	0.03	0.06	0.48	−0.19	0.07	−2.50*
Explanation	1.13	0.18	6.23***	0.23	0.06	4.10***	0.27	0.06	4.88***	0.43	0.06	7.58***	−0.58	0.07	−7.50***

$N_{Level1} = 876$, $N_{Level2} = 73$.
[a]Logistic multilevel regression analysis.
[b]Continuous multilevel regression analysis.
***$p < 0.001$, *$p < 0.05$.

archived themselves (learning group: learned and archived) vs. unfamiliar files, i.e., files that already existed before participants started working with the literature management system, had no effect on the dependent variables (see Tables 9A,B) with one exception: Participants reported less information uncertainty when the to-be-deleted files were familiar.

Hypothesis 2 proposed a mediating role of information uncertainty for the relationship between explanations and acceptance of the deletion proposal, trust, and credibility (Hypotheses 2a–c). As expected, we found that explanations reduced information uncertainty, which in turn led to increased acceptance of deletion proposals (indirect effect = 0.07, 95% CI [0.05; 0.09], $p < 0.001$), cognitive trust (indirect effect = 0.23, 95% CI [0.17; 0.30], $p < 0.001$), affective trust (indirect effect = 0.25, 95% CI [0.18; 0.32], $p < 0.001$), and credibility (indirect effect = 0.18, 95% CI [0.12; 0.23], $p < 0.001$). For the frequency of checking the explanations, a further indicator of trust, no indirect effect was found for either opening the file system (indirect effect = 0.00, 95% CI [−0.01; 0.01], $p = 0.890$) or time spent checking the file system (indirect effect = −0.03, 95% CI [−0.13; 0.07], $p = 0.569$).

Need for cognition (Hypotheses 3a–d) did not moderate the relationship between explanations and acceptance of the proposals, trust, or information uncertainty (all $zs/ts < |1.72|$, all $ps > 0.087$), but did moderate the relationship with credibility ($\gamma = -0.31$, $SE = 0.14$, $t = -2.22$, $p = 0.027$): Participants with lower (−1 *SD*) need for cognition considered the system as more credible when explanations where given (simple slope = 0.56, $t = 6.93$, $p < 0.001$). For participants with higher (+1 *SD*) need for cognition, the direction of the effect remained the same, but

TABLE 9B Experiment 3: Effects of explanations on checking the system.

	Checking the file system[a]			Time spent checking the file system[b,c]		
Predictor	***Est.***	***SE***	***z***	***Est.***	***SE***	***t***
Constant	2.41	0.77	3.15**	6.94	0.55	12.61***
Learning	1.24	1.08	1.15	0.80	0.78	1.03
Time	−0.07	0.03	−1.91	−0.11	0.03	−3.91***
Familiarity	0.46	0.24	1.92	0.30	0.19	1.63
Explanation	−0.79	0.24	−3.26**	−0.54	0.19	−2.92**

$N_{Level1} = 876$, $N_{Level2} = 73$.
[a]Logistic multilevel regression analysis.
[b]Continuous multilevel regression analysis.
[c]Logarithmic (originally in milliseconds).
***$p < 0.001$, **$p < 0.01$.

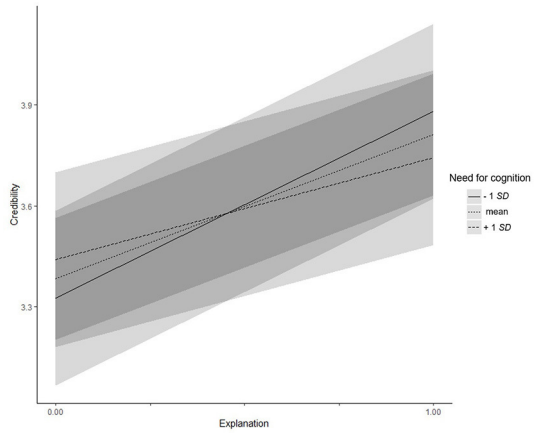

FIGURE 3
Experiment 3: Interaction effect between need for cognition and explanations on credibility. Low and high levels of need for cognition represent one standard deviation below and above the mean, respectively.

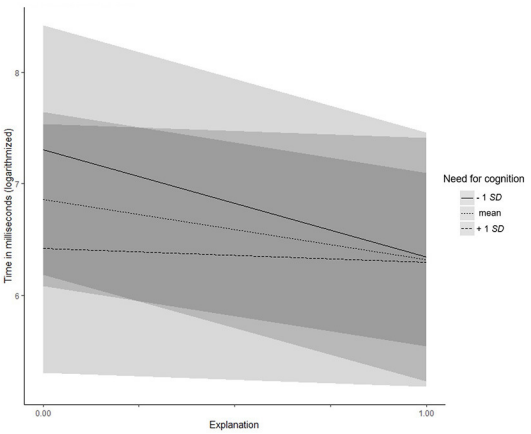

FIGURE 4
Experiment 3: Interaction effect between need for cognition and explanations on time spent checking the file system. Low and high levels of need for cognition represent one standard deviation below and above the mean, respectively.

turned out to be weaker (simple slope = 0.30, $t = 3.79$, $p < 0.001$; see Figure 3). Moreover, need for cognition moderated the effect of explanations on the time participants spent checking the file system ($\gamma = 1.05$, $SE = 0.47$, $t = 2.24$, $p = 0.025$). Participants with lower (−1 *SD*) need for cognition checked the system for a shorter time when an explanation was provided (simple slope = −0.96, $t = -3.66$, $p < 0.001$), whereas there was no difference for participants with higher (+1 *SD*) need for cognition (simple slope = −0.12, $t = -0.48$, $p = 0.632$, see Figure 4). Hypothesis 3 was partly supported.

For conscientiousness (Hypotheses 4a–d), none of the interaction effects turned out to be significant, all *z*s/*t*s < |1.90|, all *p*s > 0.059. Therefore, Hypothesis 4 was not supported.

Finally, we investigated whether the deletion of files promoted forgetting, particularly for well-known files (learning condition; Hypothesis 5). However, deleting a file was not related to recognition of file names ($\gamma = -0.17$, $SE = 0.19$, $z = -0.90$, $p = 0.368$). Therefore, Hypothesis 5 had to be rejected.

Additional analyses

As in Experiment 2, explanations led to a higher hit rate for file names ($\gamma = 0.36$, $SE = 0.16$, $z = 2.23$, $p = 0.026$). Nor did we find more forgetting (lower hit rate) of file names in the learning condition or a significant interaction between explanations and learning conditions on the hit rates for the file names.

To further explore our data, we analyzed the variables over the course of the experiment. The results revealed that over time, acceptance of the system's deletion proposals generally increased ($\gamma = 0.13$, $SE = 0.03$, $z = 4.99$, $p < 0.001$). Moreover, over time, participants trusted the system more both cognitively ($\gamma = 0.07$, $SE = 0.01$, $t = 8.79$, $p < 0.001$) and affectively ($\gamma = 0.07$, $SE = 0.01$, $t = 8.72$, $p < 0.001$), considered it more credible ($\gamma = 0.05$, $SE = 0.01$, $t = 5.84$, $p < 0.001$), felt less information uncertainty ($\gamma = -0.09$, $SE = 0.01$, $t = -7.93$, $p < 0.001$), and spent less time checking the file system ($\gamma = -0.11$, $SE = 0.03$, $t = -3.91$, $p < 0.001$).

We further found a significant interaction effect between time and explanation predicting the probability of accepting the suggestion ($\gamma = -0.33$, $SE = 0.05$, $z = -5.76$, $p < 0.001$). When participants received an explanation for the system's suggestion, the probability of acceptance was higher and did not change over time (simple slope = −0.02, $t = -0.61$, $p = 0.542$). However, when the system did not provide an explanation, in the beginning, participants had a low acceptance rate which increased over time (simple slope = 0.31, $t = 7.12$, $p < 0.001$). At the end of the experiment, explanations did not play a role for the acceptance of suggestions (see Figure 5).

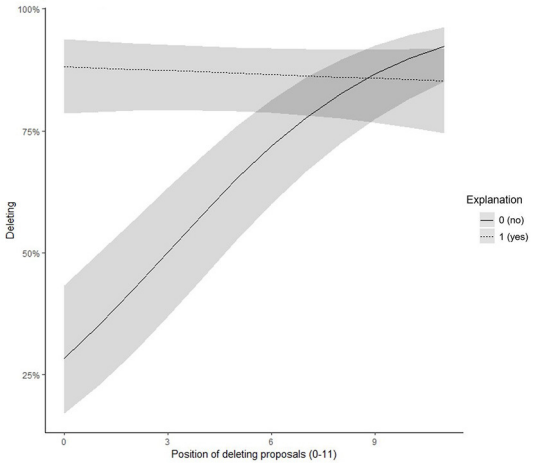

FIGURE 5
Experiment 3: Interaction effect between position of the deleting proposal and explanations on the probability of accepting the deleting proposal.

TABLE 10 Overview: Hypothesized effects of explanations on dependent variables for Experiments 1, 2, and 3.

	Experiment 1	Experiment 2	Experiment 3
Deleting file[a]	✓	✓	✓
Cognitive trust	x	x	✓
Affective trust	x	x	✓
Checking file[a,b]	–	x	✓
Credibility	x	✓	✓
Uncertainty	x	✓	✓
Recognition[a]	Poorer	Better	Better

[a]Dichotomous variable.
[b]Only assessed in Experiments 2 and 3.

In sum, the finding in Experiment 2 that explanations foster acceptance of suggestions, trust, credibility, and information uncertainty could be replicated. Moreover, the mediating role of information uncertainty was confirmed. Again, we found no effect of explanations and deletion of files on participants' memory.

Discussion

Successfully managing and deleting digital data at work becomes increasingly important. In three experiments, we investigated how individuals respond to an explanatory interactive AI system (Dare2Del), which provides suggestions to delete irrelevant digital objects. To identify important parameters for the user acceptance of these suggestions, we systematically varied several design features within and across the experiments: The presence and kind of explanations as well as the familiarity of the to-be-deleted files were varied in all experiments. In Experiments 2 and 3, we additionally provided the possibility to check the correctness of the provided explanations in the file system. In Experiment 3, we further tested whether it makes a difference if to-be-deleted files are very well known. An overview of the effects of explanations on the outcome variables for all three experiments is provided in Table 10.

Across all experiments, our findings demonstrate a general effectiveness of regularly prompting and supporting users to delete irrelevant data files, as users generally complied with these suggestions in a large number of cases, and deleted the files. Participants deleted even more files when they had the opportunity to check the appropriateness of Dare2Del's suggestions (Experiments 2 and 3). Moreover, our results also highlight the importance of providing explanations: Explanations increased the acceptance of the suggestions in all three experiments. With regard to information uncertainty, trust and credibility, it seems to play an important role whether explanations are comprehensible and can be verified in the system: If this possibility was provided, explanations also decreased information uncertainty (Experiment 2, Experiment 3), resulted in higher trust (Experiment 3) and credibility ratings (Experiment 2, Experiment 3). We assume that the absence of significant trust effects in Experiment 2 was owed to the smaller sample size, which was probably not able to detect the rather small effect.

In all experiments, the familiarity of files had no impact on the acceptance of the suggestions, information uncertainty, trust, and credibility ratings. Moreover, it did not matter whether the digital objects were well-known or not (Experiment 3). One possible explanation is that although the titles were familiar or even well-known, participants did not know details about the content of the documents. The lack of a reference to the content could therefore account for the absence of the hypothesized effects.

We also found that levels of trust, credibility, and information uncertainty changed over time. In Experiment 1, participants surprisingly showed less trust, less credibility, and more information uncertainty over time. We argue that the lack of the system's transparency (i.e., no opportunity to verify deleting suggestions) might have been responsible for these negative effects. However, in Experiment 3 (but not in Experiment 2), we found an increase in trust, credibility, and a decrease in information uncertainty over time. We also found an increase in the acceptance of the system's deleting suggestions without explanations: At the beginning, these suggestions were hardly accepted, but over time, participants gained confidence and increasingly accepted them. We assume that over time, participants were more familiar with the system, felt more confident with its handling and therefore decided to follow its suggestions more often. Whereas explanations helped to

overcome uncertainty and build trust in the beginning, they were replaced by experience and inherent trust and therefore no longer needed after a certain period of interaction time.

Information uncertainty mediated effects of explanations on accepting deleting suggestions, trust, and credibility, thus underlining its essential role for the user's acceptance of the system. In contrast, person characteristics such as need for cognition and conscientiousness seem to play a rather minor role—although we found some effects: In Experiment 2, participants with lower conscientiousness reported less information uncertainty when explanations were provided. In Experiment 3, participants with lower need for cognition considered the system as more credible and checked the system for a shorter time when explanations were given. These effects demonstrate that people low in conscientiousness and need for cognition benefit *more* from the presence of explanations: They are more likely to believe the system without questioning or wanting to thoroughly check the adequacy of the deletion suggestions. In contrast, people with high conscientiousness and a high need for cognition may not be convinced by the rather simple explanations we have given, feel more uncertain, and want to check the appropriateness of the explanation on their own (cf. Gajos and Chauncey, 2017; Ghai et al., 2021). Future research might address this issue by systematically varying the level of detail of the provided explanations.

Based on the concept of distributed cognition (Hutchins, 1995), we further investigated whether providing explanations and deleting irrelevant digital objects also promoted the forgetting of related content in human memory. Different patterns of results emerged here: Whereas the presence of explanations led to poorer recognition in Experiment 1, we even found better recognition rates in Experiments 2 and 3. We assume that this is due to the systematic design differences between the experiments: As participants had the opportunity to check the underlying file systems Experiments 2 and 3, they probably studied the file names more intensively. This higher processing depth might have strengthened subsequent recognition and thus counteracted forgetting. Interestingly, accepting deleting suggestions also led to impaired recognition in Experiment 1, but we were not able to replicate these effects in Experiments 2 and 3. The finding suggests that there might be a connection between deleting files from external storage systems and their related internal representations. However, this effect vanishes, when other factors require a deeper processing of the file names to make an informed decision as in Experiments 2 and 3.

Implications

The findings of our experiments contribute to the existing research both theoretically and practically. First, and in line with prior research (e.g., Pu and Chen, 2007; Mercado et al., 2016), our results confirm that providing explanations is an important and effective design factor, which positively influences the interaction with assistive systems and the acceptance of their suggestions. Within the present research, we also show why explanations can be beneficial to follow the system's suggestions: Explanations can reduce information uncertainty, and therefore lead to more trust and higher acceptance of the presented recommendations. These findings highlight the importance of reducing information uncertainty to enable fast and effective decisions. However, to actually be able to reduce information uncertainty, assistive systems and their suggestions need to be transparent and verifiable (see also Miller, 2018; Muggleton et al., 2018). Our experiments provide empirical evidence to previous considerations and strongly recommend considering comprehensibility and transparency in the future design of interactive AI systems. Thereby, our results suggest that the positive effects of providing explanations for deletion suggestions can even be extended by giving users the possibility to check whether the suggestions are correct. This is a novel feature which is comparably easy to implement, but could have promising effects in terms of user's acceptance and cooperation with assistive systems.

Beyond that, we found no clear evidence of immediate consequences from deleting external computer files on related internal memory representations. Although some of our results (Experiment 1) are in line with the distributed cognition approach (Hutchins, 1995) and suggested a connection between deleting and forgetting, the effect seems to be rather weak and susceptible to many context factors (e.g., memory processing depth, see below). Moreover, other cognitive phenomena such as benefits of cognitive offloading (Risko and Gilbert, 2016) or saving-enhanced memory (Storm and Stone, 2015) could come into play and prevent successful forgetting when actually saved files should be deleted. Further research is needed to elaborate how these phenomena interact, how the organization of our digital work environment and related mental representations are connected, and for whom and when deleting files can ultimately lead to a relief of human memory.

Strengths, limitations, and further research

A clear strength of the present research lies is the systematic variation of design features within and across experiments. By successively adjusting the system's parameters (i.e., providing the possibility to check suggestions), we were able to identify and elaborate the most important aspects for successful human-computer interactions. Beyond that, we also addressed possible underlying mechanisms and tested possible cross-connections between the external storage of files and related human memory.

Nevertheless, some limitations of the current studies should be noted. First, we only relied on student samples in all experiments. Although using homogeneous samples is not unusual in experimental research to keep possible disruptive factors constant, it limits the variance and the generalizability of the findings: We assume that the examined person variables (need for cognition, conscientiousness) are above average in our sample. This may have prevented the identification of more substantial individual differences.

Second, we tested our hypotheses using one specific context and task across our experiments—namely deleting files. Although this task is widely application-related as most people save (too) many irrelevant digital objects on electronic devices at work and in their private life, the question remains whether our results can be generalized to other tasks and actions. This might be especially valid for the behavioral outcome (i.e., deleting files)—but less for trust and credibility, as these outcomes have already been examined in relation to explanations in different contexts (Pu and Chen, 2007; Pieters, 2011; Shin, 2021). Moreover, in our experiments, participants worked with Dare2Del for about an hour, which may not have been enough time to get familiar with the system or to develop trust and interaction routines. Thus, our results rather reflect interaction processes when new assistive systems are introduced than long-lasting work routines. Future research should investigate the effects of explanations on affective, cognitive and behavioral outcomes for a longer time and with file systems which are much more familiar to the participants.

The third point refers to our measurement of forgetting. It may be that we failed to find a substantial forgetting effect also due to a too short time interval between the main experiment and the recognition task. We already mentioned high processing depth as possible reason for this phenomenon. However, it may be that dealing with a file—which is required to finalize the deleting decision—increases its accessibility in the short term. Nevertheless, it might still help to detach from and forget it in the long term. Therefore, it would be interesting to explore long-term memory effects in future studies. In addition, we used a recognition task. Research on intentional forgetting has shown that forgetting effects are stronger and consistent in free recall tests but not compulsorily in recognition tests (MacLeod, 1975). Thus, future experiments might use also free recall to investigate forgetting.

Conclusion

The present study examined interactions between humans and interactive computer systems supporting users to delete irrelevant data files. Results underlined the importance of presenting explanations for the acceptance of deleting suggestions, but also point to the need of their transparency and verifiability to generate trust. However, we did not find clear evidence for immediate cross-connections between deleting computer files and human forgetting of the related mental representations.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/dk6en/.

Ethics statement

The studies involving human participants were reviewed and approved by Deutsche Gesellschaft für Psychologie. The patients/participants provided their written informed consent to participate in this study.

Author contributions

KG and CN were responsible for study design and wrote the manuscript. KG conducted the data collection and analyzed the data. SS provided technical support for the experimental setup. US contributed to the study design and supported writing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The project (NI 1066/4-1) was funded by the German Research Council (Deutsche Forschungsgemeinschaft, DFG).

Acknowledgments

We thank Ramona Crämer, Christoph Gründinger, Ann-Paulin Nutz, and Sophie Zech for supporting data collection, and Angelina Olivia Wilczewski for contributing to the current Dare2Del prototype.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those

of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.919534/full#supplementary-material

References

Anderson, M. C., and Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends Cogn. Sci.* 18, 279–292. doi: 10.1016/j.tics.2014.03.002

Arend, M. G., and Schäfer, T. (2019). Statistical power in two-level models: a tutorial based on Monte Carlo simulation. *Psychol. Methods* 24, 1–19. doi: 10.1037/met0000195

Asterhan, C. S., and Schwarz, B. B. (2009). Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. *Cogn. Sci.* 33, 374–400. doi: 10.1111/j.1551-6709.2009.01017.x

Barrick, M. R., and Mount, M. K. (1991). The Big Five personality dimensions and job performance: a meta-analysis. *Pers. Psychol.* 44, 1–26. doi: 10.1111/j.1744-6570.1991.tb00688.x

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Berenbaum, H., Bredemeier, K., and Thompson, R. J. (2008). Intolerance of uncertainty: exploring its dimensionality and associations with need for cognitive closure, psychopathology, and personality. *J. Anxiety Disord.* 22, 117–125. doi: 10.1016/j.janxdis.2007.01.004

Biran, O., and Cotton, C. (2017). "Explanation and justification in machine learning: A survey," in *IJCAI-17 Workshop on Explainable AI (XAI)*, Vol. 8. p. 8–13. Available online at: https://scholar.google.com/scholar_lookup?author=O.+Biran&author=C.+Cotton+&publication_year=2017&title=%E2%80%9CExplanation+and+justification+in+machine+learning%3A+a+survey,%E2%80%9D&journal=IJCAI-17+Workshop+on+Explainable+AI+(XAI)&volume=Vol.%208&pages=1

Bjork, E. L., Bjork, R. A., and Anderson, M. C. (1998). "Varieties of goal-directed forgetting," in *Intentional Forgetting: Interdisciplinary Approaches*, eds J. M. Golding, and C. M. MacLeod (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 103–137.

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., and Schwarz, N. (1994). Need for cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Need for cognition: a scale measuring engagement and happiness in cognitive tasks]. *Zeitschrift für Sozialpsychologie* 25, 147–154.

Bruckert, S., Finzel, B., and Schmid, U. (2020). The next generation of medical decision support: a roadmap toward transparent expert companions. *Front. Artif. Intellig.* 3, 507973. doi: 10.3389/frai.2020.507973

Cacioppo, J. T., and Petty, R. E. (1982). The need for cognition. *J. Pers. Soc. Psychol.* 42, 116–131. doi: 10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., and Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychol. Bull.* 119, 197–253. doi: 10.1037/0033-2909.119.2.197

Chen, C., Liu, C., Huang, R., Cheng, D., Wu, H., Xu, P., et al. (2012). Suppression of aversive memories associates with changes in early and late stages of neurocognitive processing. *Neuropsychologia* 50, 2839–2848. doi: 10.1016/j.neuropsychologia.2012.08.004

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edn. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Costa, P. T., and McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *J. Pers. Disord.* 6, 343–359. doi: 10.1521/pedi.1992.6.4.343

Costa, P. T., McCrae, R. R., and Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: a revision of the NEO personality inventory. *Pers. Individ. Dif.* 12, 887–898. doi: 10.1016/0191-8869(91)90177-D

Dabbish, L. A., and Kraut, R. E. (2010). Email overload at work: an analysis of factors associated with email strain. *IEEE Eng. Manage. Rev.* 38, 76–90. doi: 10.1109/EMR.2010.5494697

Edmunds, A., and Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *Int. J. Inf. Manage.* 20, 17–28. doi: 10.1016/S0268-4012(99)00051-1

Fogg, B. J., Swani, P., Treinen, M., Marshall, J., Laraki, O., Osipovich, A., et al. (2001). "What makes web sites credible? A report on a large quantitative study," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '01*, 61–68.

Foster, N. L., and Sahakyan, L. (2011). The role of forget-cue salience in list-method directed forgetting. *Memory* 19, 110–117. doi: 10.1080/09658211.2010.537665

Gajos, K. Z., and Chauncey, K. (2017). "The influence of personality traits and cognitive load on the use of adaptive user interfaces," in *Proceedings of the 22Nd International Conference on Intelligent User Interfaces, IUI '17*, 301–306.

Ghai, B., Liao, Q. V., Zhang, Y., Bellany, R., and Mueller, K. (2021). Explainable active learning (XAL): toward AI explanations as interfaces for machine teachers. *Proc. ACM Hum. Comp. Interact.* 4, 235. doi: 10.1145/3432934

Göbel, K., and Niessen, C. (2021). Thought control in daily working life: how the ability to stop thoughts protects self-esteem. *Appl. Cogn. Psychol.* 35, 1011–1022. doi: 10.1002/acp.3830

Hair, M., Renaud, K. V., and Ramsay, J. (2007). The influence of self-esteem and locus of control on perceived email-related stress. *Comput. Human Behav.* 23, 2791–2803. doi: 10.1016/j.chb.2006.05.005

Haynes, W. (2013). "Bonferroni correction," in *Encyclopedia of Systems Biology*, eds W. Dubitzky, O. Wolkenhauer, K. H. Cho, and H. Yokota (New York, NY: Springer).

Henkel, L. A. (2014). Point-and-shoot memories: the influence of taking photos on memory for a museum tour. *Psychol. Sci.* 25, 396–402. doi: 10.1177/0956797613504438

Hoskins, A. (2016). Memory ecologies. *Memory Stud.* 9, 348–357. doi: 10.1177/1750698016645274

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: The MIT Press.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* 15, 309–334. doi: 10.1037/a0020761

Keil, F. (2006). Explanation and understanding. *Annu. Rev. Psychol.* 57, 227–254. doi: 10.1146/annurev.psych.57.102904.190100

Kluge, A., and Gronau, N. (2018). Intentional forgetting in organizations: the importance of eliminating retrieval cues for implementing new routines. *Front. Psychol.* 9, 51. doi: 10.3389/fpsyg.2018.00051

Kofta, M., and Sedek, G. (1999). Uncontrollability as irreducible uncertainty. *Eur. J. Soc. Psychol.* 29, 577–590. doi: 10.1002/(SICI)1099-0992(199908/09)29:5/6<577::AID-EJSP947>3.0.CO;2-K

Komiak, S. X., and Benbasat, I. (2004). Understanding customer trust in agent-mediated electronic commerce, web-mediated electronic commerce, and traditional commerce. *Inform. Technol. Manage.* 5, 181–207. doi: 10.1023/B:ITEM.0000008081.55563.d4

Komiak, S. Y., and Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *Manage. Inform. Syst. Q.* 30, 941 – 960. doi: 10.2307/25148760

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* 10, 1–8. doi: 10.1038/s41467-019-08987-4

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends Cogn. Sci.* 20, 748–759. doi: 10.1016/j.tics.2016.08.001

MacLeod, C. M. (1975). Long-term recognition and recall following directed forgetting. *J. Exp. Psychol. Hum. Learn. Memory* 1, 271–279. doi: 10.1037/0278-7393.1.3.271

Mathieu, J. E., Aguinis, H., Culpepper, S. A., and Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *J. Appl. Psychol.* 97, 951–966. doi: 10.1037/a0028380

Meeßen, S. M., Thielsch, M. T., and Hertel, G. (2020). Trust in management information systems (MIS): a theoretical model. *Zeitschrift für Arbeits und Organisationspsychologie A O* 64, 6–16. doi: 10.1026/0932-4089/a000306

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., and Procci, K. (2016). Intelligent agent transparency in human-agent teaming for multi-uxv management. *Hum. Factors* 58, 401–415. doi: 10.1177/0018720815621206

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., and Besold, T. (2018). Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Mach. Learn.* 107, 1119–1140. doi: 10.1007/s10994-018-5707-3

Müller, L. S., Meeßen, S. M., Thielsch, M. T., Nohe, C., Riehle, D. M., and Hertel, G. (2020). "Do not disturb! Trust in decision support systems improves work outcomes under certain conditions," in *Mensch und Computer 2020 – Tagungsband*, eds F. Alt, S. Schneegass, and E. Hornecker (New York: ACM), 229–237.

Nezlek, J. B. (2012). "Multilevel modeling for psychologists," in *APA Handbook of Research Methods in Psychology, Vol. 3. Data Analysis and Research Publication*, eds H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher (Washington, DC: American Psychological Association), 219–241.

Niessen, C., Göbel, K., Lang, J., and Schmid, U. (2020a). Stop thinking: an experience sampling study on suppressing distractive thoughts at work. *Front. Psychol.* 11, 1616. doi: 10.3389/fpsyg.2020.01616

Niessen, C., Göbel, K., Siebers, M., and Schmid, U. (2020b). Time to forget: intentional forgetting in the digital world of work. *German J. Work Org. Psychol.* 64, 30–45. doi: 10.1026/0932-4089/a000308

Ostendorf, F., and Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R; Manual Revidierte Fassung*. Göttingen: Hogrefe.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychol. Methods* 19, 459–481. doi: 10.1037/a0036434

Perry, M. (2003). "Distributed cognition," in *Interactive Technologies, HCI Models, Theories, and Frameworks*, ed J. M. Carroll (San Francisco, CA: Morgan Kaufmann), 193–223.

Pieters, W. (2011). Explanation and trust: what to tell the user in security and AI? *Ethics Inf. Technol.* 13, 53–64. doi: 10.1007/s10676-010-9253-3

Pu, P., and Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge Based Syst.* 20, 542–556. doi: 10.1016/j.knosys.2007.04.004

Randall, J. G., Oswald, F. L., and Beier, M. E. (2014). Mind-wandering, cognition, and performance: a theory-driven meta-analysis of attention regulation. *Psychol. Bull.* 140, 1411–1431. doi: 10.1037/a0037428

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you? Explaining the predictions of any classifier," in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM), 1135–1144.

Risko, E. F., and Gilbert, S. J. (2016). Cognitive offloading. *Trends Cogn. Sci.* 20, 676–688. doi: 10.1016/j.tics.2016.07.002

Rong, H., Zhang, H., Xiao, S., Li, C., and Hu, C. (2016). Optimizing energy consumption for data centers. *Renewable Sustain. Energy Rev.* 58, 674–691. doi: 10.1016/j.rser.2015.12.283

Sahakyan, L., Delaney, P. F., and Goodmon, L. B. (2008). Oh, honey, I already forgot that: strategic control of directed forgetting in older and younger adults. *Psychol. Aging* 23, 621–633. doi: 10.1037/a0012766

Schmid, U. (2021). "Interactive learning with mutual explanations in relational domains," in *Human-Like Machine Intelligence*, eds S. Muggleton, and N. Chater (Oxford: Oxford University Press), 338–354.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *Int. J. Hum. Comput. Stud.* 146, 102551. doi: 10.1016/j.ijhcs.2020.102551

Soares, J. S., and Storm, B. C. (2022). Does taking multiple photos lead to a photo-taking-impairment effect? *Psychon. Bull. Rev*. doi: 10.3758/s13423-022-02149-2

Soucek, R., and Moser, K. (2010). Coping with information overload in email communication: evaluation of a training intervention. *Comput. Human Behav.* 26, 1458–1466. doi: 10.1016/j.chb.2010.04.024

Sparrow, B., Liu, J., and Wegner, D. M. (2011). Google effects on memory: cognitive consequences of having information at our fingertips. *Science* 333, 776–778. doi: 10.1126/science.1207745

Storm, B. C., and Stone, S. M. (2015). Saving-enhanced memory: the benefits of saving on the learning and remembering of new information. *Psychol. Sci.* 26, 182–188. doi: 10.1177/0956797614559285

Sutton, J. (2016). "Scaffolding memory: themes, taxonomies, puzzles," in *Contextualizing Human Memory: An Interdisciplinary Approach to Understanding How Individuals and Groups Remember the Past*, eds C. Stone and L. Bietti (Abingdon: Routledge/Taylor & Francis Group), 187–205.

Thaler, A., and Schmid, U. (2021). "Explaining machine learned relational concepts in visual domains-effects of perceived accuracy on joint performance and trust," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 43, 1705–1711.

Thielsch, M. T., Meeßen, S. M., and Hertel, G. (2018). Trust and distrust in information systems at the workplace. *PeerJ* 6, e5483. doi: 10.7717/peerj.5483

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). Mediation: R package for causal mediation analysis. *J. Stat. Softw.* 59, 1–38. doi: 10.18637/jss.v059.i05

Tintarev, N., and Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Model. User Adapt. Interact.* 22, 399–439. doi: 10.1007/s11257-011-9117-5

Van den Bos, K. (2009). Making sense of life: The existential self trying to deal with personal uncertainty. *Psychol. Inq.* 20, 197–217. doi: 10.1080/10478400903333411

Wang, W., and Benbasat, I. (2007). Recommendation agents for electronic commerce: effects of explanation facilities on trusting beliefs. *J. Manage. Inform. Syst.* 23, 217–246. doi: 10.2753/MIS0742-1222230410

Wilson, T. D., Centerbar, D. B., Kermer, D. A., and Gilbert, D. T. (2005). The pleasures of uncertainty: prolonging positive moods in ways people do not anticipate. *J. Pers. Soc. Psychol.* 88, 5–21. doi: 10.1037/0022-3514.88.1.5

Zhang, J., and Patel, V. L. (2006). Distributed cognition, representation, and affordance. *Pragmat. Cogn.* 14, 333–341. doi: 10.1075/pc.14.2.12zha

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 07 February 2023
DOI 10.3389/frai.2022.960384

OPEN ACCESS

EDITED BY
Dietrich Albert,
University of Graz, Austria

REVIEWED BY
Witold M. Wachowski,
Marie Curie-Sklodowska University, Poland
Hans Van Eyghen,
VU Amsterdam, Netherlands

*CORRESPONDENCE
Rasmus Gahrn-Andersen
✉ rga@sdu.dk

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 02 June 2022
ACCEPTED 30 December 2022
PUBLISHED 07 February 2023

CITATION
Cowley SJ and Gahrn-Andersen R (2023) How systemic cognition enables epistemic engineering. *Front. Artif. Intell.* 5:960384. doi: 10.3389/frai.2022.960384

How systemic cognition enables epistemic engineering

Stephen J. Cowley and Rasmus Gahrn-Andersen*

Department of Language and Communication, University of Southern Denmark, Slagelse, Denmark

Epistemic engineering arises as systems and their parts develop functionality that is construed as valid knowledge. By hypothesis, epistemic engineering is a basic evolutionary principle. It ensures that not only living systems identify the differences that make differences but also ensure that distributed control enables them to *construct* epistemic change. In tracking such outcomes in human life, we stress that humans act within poly-centered, distributed systems. Similar to how people can act as inert parts of a system, they also actively bring forth intents and vicariant effects. Human cognitive agents use the systemic function to construct epistemic novelties. In the illustration, we used a published experimental study of a cyborg cockroach to consider how an evoneered system enables a human subject to perform as an adaptor with some "thought control" over the animal. Within a wide system, brains enable the techniques to arise *ex novo* as they attune to the dictates of a device. Human parts act as adaptors that simplify the task. In scaling up, we turn to a case of organizational cognition. We track how adaptor functions spread when drone-based data are brought to the maintenance department of a Danish utility company. While pivoting on how system operators combine experience with the use of software, their expertise sets off epistemically engineered results across the company and beyond. Vicariant effects emerge under the poly-centered control of brains, persons, equipment, and institutional wholes. As a part of culture, epistemic engineering works by reducing entropy.

KEYWORDS

distributed cognition, social organizing, simplexity, systemic cognition, radical embodied cognitive science, pre-reflective experience, vicariance, evoneering

1. Introduction

In Europe and America, knowing is often ascribed to an organism, body, mind, or brain. In contrast to, say, Chinese or African traditions, the individual is treated as the locus of both know-how and reason. In making a link between anthropology and computational models, Hutchins (1996) brings new light to how collective knowing enables to inform human agency. In allowing cognitive distribution, he traces epistemic outcomes across systems that lack a single locus of control. When rowing canoes across the Pacific or, indeed, bringing a ship into port, people link up beliefs, devices, observations, and acting within culturally distributed systems. Knowing includes—but is not generated by—individual actors. In applying the view to science, Giere (2004) invokes how the Hubble spacecraft enabled distributed systems to bring forth new knowledge of the universe. Like other organized knowledge, poly-centered systems enable science to arise through what Giere (2004) calls "human cognitive agents." In what follows, we radicalize such views by tracking how wide systems can affect the epistemic agency of living human beings.

Primate intelligence is predominantly social (Jolly, 1966; Humphrey, 1976) and, in the last million years or so, hominins and eco-systems have co-evolved (Sterelny, 2007). Bodies and, especially, brains have brought humans the extreme plasticity that sustains practices such as trade (Ross, 2012). In Hutchins's (1996) terms, practices inform the distributed cognitive systems that link artifacts, language, and ways of acting. Hence, they include what Malafouris (2013, 2019) calls *material engagement*: in using materials such as clay, we draw upon cultural resources such

as norms and conventions as bodily promptings enable us to use techniques, skills, and methods. For Malafouris, "enactive signification" arises as parameters co-function to nudge a person to substitute one way of acting with another. Humans gain flexibility and construct epistemic powers as they actualize social practices. They perform roles and develop styles that create diversity that uses a trick of *vicariance* or how one can "perform the same tasks with different systems, solutions or behaviors" (Berthoz and Tramus, 2015, p. 1–2). Crucially, since vicariance serves bodies, brains, and social activities (Cowley and Gahrn-Andersen, 2022), it creates novelty by reducing entropy or uncertainty (usually, if not always, by changing the parameters of a system). Vicariant effects spread across bodily modalities, social groups, and neural organization and, as a result, parties gain as epistemic change self-fabricates within cognitive systems.

In pursuing how such vicariant effects are brought about, the article begins with a "minimal" case. We describe how, in an experimental setting, a system sets off epistemic change as a person comes to exert "thought control" over a cockroach. Agency links an engineered system, human-cockroach interdependencies, pre-reflective experience, and a brain that constructs and sustains bodily techniques. Highlighting the systemic, we emphasize how the human adaptor uses cognition beyond the body. Later, we compare the neural parameter setting of the cockroach experiment to how vicariant effects spread when drones were introduced to a Danish utility company. In both cases, people reduce entropy (uncertainty) within wide cognitive systems as, often without knowing why, they set off effects that serve a wider system: vicariant outcomes thus transform both individual performance and the company task regime.

2. Cognition—The role of "knowledge" for systems

A distributed perspective on cognition (Hutchins, 1996; Rogers, 1997; Perry, 2013) first emerged as a counterpoint to core tenets of orthodox cognitivism (e.g., Fodor, 1975; Marr, 1982; Searle, 1992). It does so in that the classic cognitive view treats the organism as the "source" of intelligent behavior. In philosophical guise, knowledge is ascribed to sense impressions, mind, and reason; by contrast, with cognitive science, attention falls on learning, computation, sense-making, organism-environment coupling, etc. Turning to working environments, Hutchins showed that, in many cases, such models are demonstrably inadequate. There is no organismic source of cognition in, say, navigation. Rather, people incontrovertibly draw on cultural resources and wide systems (Wilson, 2004) to achieve epistemic outcomes. Socially organized activity is a dynamical interplay of agents and environments which link cognitive practices with, above all technologies and external representation media. In a distributed system, social practices or organizations sustain heterogeneous kinds of processes. The distributed perspective thus applies to practices as diverse as, say, crime scene investigation (Baber, 2010), medical situation awareness (Fioratou et al., 2016), insight problem-solving (Vallée-Tourangeau and Wrightman, 2010), or, indeed, how a daughter decorously tries to quieten her mother (Cowley, 2014).

The entire cognitive system unites a myriad of parts as "inner and external" resources co-function in diverse ways (cf. Michaelian and Sutton, 2013, p. 10). As Hutchins (2014) came to phrase it in theoretically oriented work, the perspective applies to all of human cognition: it characterizes "the microprocesses of interaction across the diverse components of these distributed and heterogeneous cognitive systems" (Hutchins, 2014, p. 5). Yet, as Hutchins notes, his own early work views "cognitive processes in terms of the propagation and transformation of representations" (Hutchins, 2001, p. 2068). Hence, proponents of the distributed perspective who retain a traditional model of representations find themselves committed to the "source" view of orthodox cognitivism (for a criticism, see Hutto et al., 2014). Placing intent in the brain, they treat cognizers as parties that propagate and transform "particular representational states across distinct (internal and external) media" (Michaelian and Sutton, 2013, p. 5). Whereas, Hutchins began with a focus on representations in a literal sense (Hutchins, 1996, p. 363–364), he later shifts to a more liberal view. Hence, far from addressing the role of living agency in cognition – or how intent arises – later work (Hutchins, 2020) still focuses on how externalized resources extend how people act as they perform social roles and rely on interactions. He explicitly suggests that "distributed cognition is not a kind of cognition at all, it is a perspective on cognition." His concern is with, not explaining cognition or the role of bodies in epistemic change, but, rather, how "participants to an interaction coinhabit a shared environment" (2020, p. 375). Very plausibly, Hutchins adopts the view that "interaction is the basis for the distribution of cognitive labor" (2020, p. 377). As an ethnographer, albeit an unorthodox one, he approaches people as social actors. Leaving aside issues of intent, he can overlook *how* agency changes and, on methodological grounds, changes in cultural operations. Since he asks how participants contribute to procedures, he reduces language and agency to their role in task performance. Others are more concerned with individual responsibility (Jones, 2013) or how looser systems depend on language, knowledge, and expertise (Perry, 2013). In seeking to deal with the tension, Baber et al. (2014), for example, use the concept of "affordances" to allow individual control of tools within a "person–environment–tool–object system" (p. 10). Adopting Turvey's (1992) view of affordance (Gibson, 1979), Baber et al. allow for individual expertise in control:

> Even if there are regions that are active under specific conditions, the skill of the expert tool user comes from the ability to control their activity with sufficient spare capacity to cope with future demands and to respond to the changing context in which they are using the tools to effect changes in the object being worked on. (Baber et al., 2014, p. 12)

In making individual skills and expertise partly constitutive of distributed processes, Baber et al. identified the collective-individual tension that runs through research on distributed cognition. The focus on outcomes can lead one to highlight, not individual doings, but a collective effort. For instance, Hutchins reports on how the crew of the USS Palau dealt with the issues relating to the loss of main steam (Hutchins, 1996). He traces the outcomes to how tightly coupled practices are structured around the well-understood/defined task of managing how the vessel is brought to anchor. Hutchins writes:

> The safe arrival of the Palau at anchor was due in large part to the exceptional seamanship of the bridge crew, especially the navigator. But no single individual on the bridge

> acting alone–neither the captain nor the navigator nor the quartermaster chief supervising the navigation team–could have kept control of the ship and brought it safely to anchor. (p. 5)

Although Hutchins (1996) recognizes the seamanship of the navigator, his ethnography of the supra-entity highlights interaction and participant roles. Hence, Hutchins plays down individuals, intents and propensities, how skills arise, or how they are selected. This is because, in a task context, the right choices are simply assumed. Furthermore, it is by treating a person as a social actor (not a source of cognizing) that the distributed perspective breaks with classic views. Later, we show how it allows emphasis on autonomy to be replaced by a view of agency as using poly-centered and diachronic control. Indeed, even on a standard view, this is implied where a system:

> dynamically reconfigures itself to bring subsystems into functional coordination. Many of the subsystems lie outside individual minds; in distributed cognition, interactions between people as they work with external resources are as important as the processes of individual cognition (Lintern, 2007, p. 398).

Control arises as the system co-configures its functions such that tasks are successfully accomplished. Classically, it uses extant equipment, routines, procedures, etc. or, as for Latour (2007), human and non-human parts to serve as actors ("actants"). In what follows, unlike Latour and Hutchins, we will turn to how living human bodies function as parts of wide systems.

Starting with social actors allows a single "level of analysis" to apply to organizations, practices, and ways of acting. Turning from control, Hutchins (1996) identifies distributed cognition with tightly coupled practices that, in later work (Hutchins, 2014, 2020), are explicitly said to ground all of human cognition. He uses what Cheon (2014) calls a "task-specification requirement" where activity is "distributed" around a clearly specified and collectively understood task. Such a view is exemplified by the malfunction in the steam whistle where, for the crew, their task becomes that of finding a functional substitute or vicariant solution to warning an approaching sailboat of possible collision (Hutchins, 1996, p. 4). As in Marr's (1982) work on vision, a cognitive task is computationally defined and, given formal description, separated from a (presumed) implementational level. Even Hutchins (2014) retains this view in recent work on the details of cockpit control: here too, he leaves aside implementation to focus on actions: thus, in Weibel et al. (2012), the use of eye-tracking data is reported. However, it serves to pursue, for example, the meaning of the pilot's "light touching of the front edge of left thrust lever with the side of the pinky finger on his right hand, bumping it lightly in the direction of reduced thrust" (p. 112). For methodological reasons, as Gahrn-Andersen (2021) shows, the object of study concerns how humans act as parts of well-defined cognitive systems. In other words, given an extant epistemic definition of the task, the whole system (e.g., practice, organization) is viewed as a stable, supervening entity. Control draws on predictable functionality to ensure that what is described *counts as valid knowledge*. Yet, a high price is paid by starting with a systemic whole. Human individuals become social operators in unchanging systems. Thus, for Afeltowicz and Wachowski (2015), the approach fails to qualify as a cognitive theory because it cannot clarify how intent arises. Of course, the perspective has no such goal. However, recognition of the flaw points to the interdependency of living and non-living systems. This is prefigured by Giere (2004) who, taking the distributed perspective to science, carefully distinguishes the human cognitive agent from the whole system. Without this move, one risks assuming, with Michaelian and Sutton (2013) that "expertise is not a property of individual agents, but is built in to the constraints of the system" (Michaelian and Sutton, 2013, p. 5). Not only does one leave aside how intent emerges but also one replaces a whole system's pre-established structures and loci of control (e.g., routines) with attention to operational shifts, systemic change, expertise and the entangled, and highly variable workings of living human bodies. While their functions indeed reach beyond the sum of its parts determining proper actions, only attention to a "person-in-the-system" (Fester-Seeger, 2021) can open up how systems generate intent or use vicariant effects to achieve epistemic change.

Hutchins (2014) applies his perspective to all of human cognition by comparison to the theory of extended mind. Hence, task-based human cognition falls within the constraints of "cultural eco-systems." He views how agents perform –act, draw, and speak – as "participants" in wider systems: hence as in earlier work, his focus is collective. Indeed, an ecosystemic focus abstracts away from actual doings and organized action. Hutchins seek to "shift the focus from ecological assemblies surrounding an individual person to cultural ecosystems operating at larger spatial and temporal scales" (2014, p. 35). Of course, at a descriptive level, he recognizes that individual participant matters (e.g., as in the case of a flight crew's visual attention which is structurally determined by the practice of preparing for descent 2014, p. 44). Theoretically, however, he emphasizes systemic stability or how existing practices are sustained. In his terms, "the stability, resilience, or persistence of a practice depends on the network of relations to other practices within which it is embedded" (p. 46). Indeed, Hutchins emphasizes a "web of cultural regularities" and, with these, the cultural practices, which sustain them (2014, p. 47). As he notes, the perspective allows practices to reduce contingencies to the extent that those familiar with a relevant ecosystem will experience similar phenomena as belonging to the same type (e.g., perceiving a line of people as a queue). Importantly, he notes how "cultural practices decrease entropy and increase the predictability of experience" (2014, p. 46). In this context, even individual learning is structurally determined by ecosystemic regularities. The perspective thus treats both individual and collective experiences as intrinsic to the operations that guarantee systemic reproduction. By implication, parts (e.g., workers or equipment) and procedures are functionally replaceable. This takes us back to our criticism of Hutchins (1996): By taking the supra-entity as given-in-advance, he fails to interrogate how epistemic shifts occur. Rather, his system is functionally indifferent to the substitution of its elements and actual ways of performance. Instead of exploring intents, systemic adjustment, change, and development, vicariance is separated from persons and systemic dysfunction or, indeed, significant operational change.

While a truism that human agency and power are socially distributed, we turn to how parameters operate as events arise in epistemic domains. Building on viewing language as distributed by how embodiment informs agency (Blair and Cowley, 2003; Cowley, 2011, 2014), we highlight systemic interdependency. Similar to what Giere (2004) shows for science or Vallée-Tourangeau and Wrightman (2010) for individual differences in mental arithmetic, we stress that persons are interdependent with non-living parts of wider systems.

As illustrated below, these prompt epistemic change in, at times, neural organization and, at others, an organized task regime. A wider system induces vicariant effects as persons engage with things and each other. Each person-in-the system is a social actor (i.e., a living being and a participant) who contributes to cascading systemic change (in various scales). Often, epistemic change is triggered as an agent draws on what appears as an *ex novo* event. Turning to functional coordination and stability, we stress how distributed agency (refer to Enfield, 2013) drives epistemic change. Since this has a biosocial basis, human cognition links distributed systems to living bodies, language-activity (or languaging) and semiotic assemblages (Pennycook, 2017). In order to clarify how vicariant effects arise, we bring systemic ethnography to how, in actual cases, practices unfold. We unleash the power of tracing living human agency to how bodies (and brains) contribute as parts of wide systems. Individual agents draw on their embedding in larger wholes to shape traits a person's competencies (in the system). Hence, distributed parts enable organic and organized parameter setting as systemic function draws on what we call *epistemic engineering*. As a result, the process enables humans to use ecosocial resources in a life history of epistemic change. Coming to know this implicates routine performance that unites separable systems, various control centers (e.g., brains and computers) and modes of action.

As will be explained in section 5, our account turns from a computational (or supra-entity) level by treating human cognition as systemic and poly-centric. Accordingly, we play down pre-determined cognitive tasks and views that ascribe cognition to a single implementational source (i.e., a strictly autonomous system). Before turning to our systemic frame (Cowley and Vallée-Tourangeau, 2013, 2017; Secchi and Cowley, 2021; Secchi et al., 2023), we present two case studies of epistemic engineering. These illustrate (a) how cognitive systems require changing the loci of control and (b) how agents, in their capacity as such, draw on vicariant effects to affect the outcome of distributed systems.

3. The minimal case

A principle of neural re-use (Anderson, 2010) permits brains to use a body's life history as they construct bodies that develop as effective performers and, indeed, participants in distributed systems. Hence, we begin with how neural flexibility enables a person to adapt to what we call a *minimal engineered system*. Similar vicariant effects occur with, say, sensory substitution (Froese and Ortiz-Garin, 2020) or "thought" control of a prothesis (e.g., Edelman et al., 2019). While evoneered technology is often studied as of value in itself, less weight has hitherto been placed on the biotech interface or how a living brain adapts to a device. In that the results demand learned adaptation, we extend work published elsewhere (Gahrn-Andersen and Prinz, 2021) to highlight *natural* evoneering.

In the case of the cyborg cockroach, "thoughts" come to influence an insect's movements (Li and Zhang, 2016). Of course, this is not literally a matter of "thinking": rather, without knowing what he or she is doing, a person manages input to the visual cortex that is monitored by an EEG device. Since this transmits to the cockroach's antenna nerve, it sets off vicariant effects. Since a result, the cockroach comes to resemble a cyborg in that it moves, to an extent, under human control. The person gains a new way of acting: he or she uses an engineered interface within a poly-centered system. As a person-adaptor controls EEG response to a moving cockroach on a flickering screen, the subject wills "thoughts" or, more precisely, generates micro-electronic input. The subject learns to will left and right movements by influencing the cockroach's antennae nerves. Building on work which showed that cockroach moves can be shaped by radio transmission of joystick manipulation (Latif and Bozkurt, 2012), Li and Zhang (2016) added the brain-to-brain interface between EEG-output and antennae nerves. In what follows, we report on an experimental study that involved three subjects and three cockroaches. This vicariant enabling device allowed subjects to learn to use "watching and willing" to nudge a moving cockroach on an S-shaped track (refer to Figure 2). In Figure 1, we present an engineering view of the poly-centered system.

While acting as a supra-system, experimenters merely offer instructions and minimal training. Though part of the whole system, they have no active role in "looking-and-willing" or thought control. Thus, in the terms of Lintern (2007), one can ask how the whole "dynamically reconfigures itself" (p. 398). In so doing, we focus on how epistemic change arises as a subject gains some control over the cockroach. In such a case, systems and parts enable vicariant effects as a subject masters what we call a *technique*. In this "minimal" epistemic engineering, the subject (and the brain) connect: (a) how a person *assesses/manages* watching-and-willing and, thus, the adaptor's EEG output[1] and (b) how input to the antennas' nerves affects cockroach movements. If successful, the poly-centered system achieves "functional coordination" between looking, neural activity, the engineered adaptor, and the cockroach. In Lintern's (2007) terms, "external resources are as important as the processes of individual cognition" (ibid).

In producing EEG output *for* the cockroach, a human subject assesses cockroach moves while willing changes in cockroach movements (refer to Figure 2). Hence, adaptors and "thoughts" (or EEG measures) come to anticipate cockroach activity. Given repetition and experience, the human gains techniques: in an enlanguaged world, participants grasp the following: (1) what the task is; and (2) what has to be done. However, since one cannot know (in advance) what it is like to move a cyborg cockroach, techniques can only arise *ex novo*. Even if much depends on what we call skills (and can be described by theories like predictive processing), the vicariant effects do not *reduce* to brain side process. It is only as part of a brain-in-a-wide (or poly-centered) system that an engineered system can use a "composite device" constituted by the setting (and, ultimately, the work of the experimenters). In time, the accomplished use of the device and cockroach brings "synergism and functionality" to the person (Gahrn-Andersen and Prinz, 2021) who performs the experiment. Far from reducing to learning, one gains epistemic power (know-how) that is entirely dependent on the whole system: one draws on interdependencies (and repetition) in coming to act with a new kind of intent.

1 The system measures "steady state visual evoked potential" as EEG response from the visual cortex that arises in looking at the moving cockroach on a flickering screen. The EEG system is adjusted to focus on a certain bandwidth. Hence, what we call "looking and willing" involves a range of factors and, as with any such system, there are issues of noise. Thus, while subjects are asked to keep their heads still, even in the demonstration video, they track the cockroach movement in ways that are highly visible.

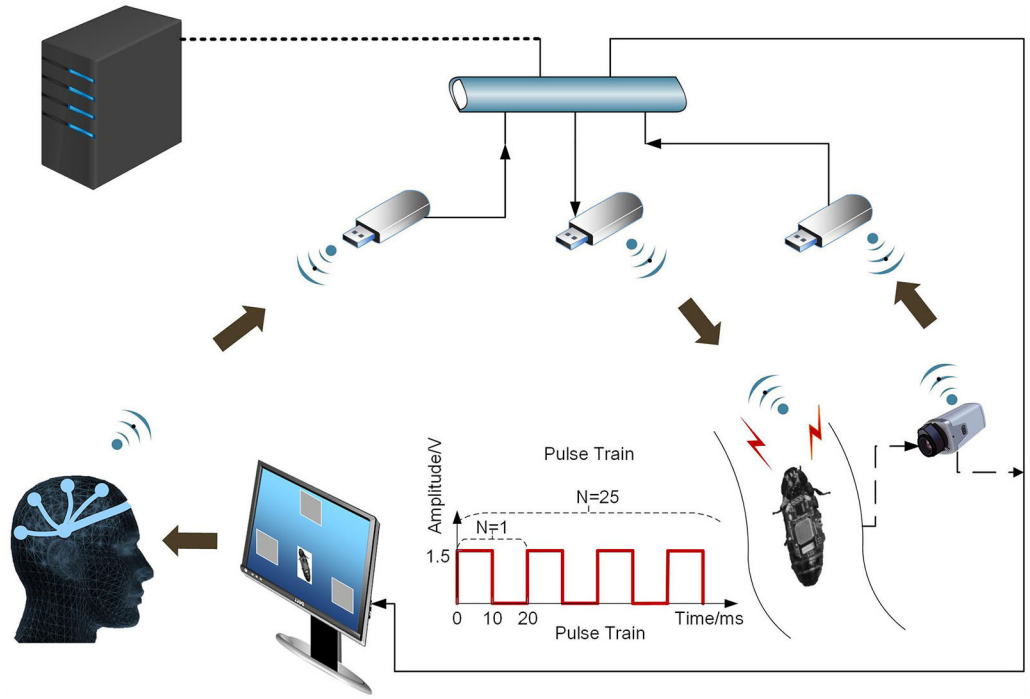

FIGURE 1
The experimental system encompasses a person, a cockroach and a steady-state visual evoked potential (SSVEP) based brain-computer interface (BCI). The distributed system works as the subject sends real time BCI commands to the cockroach as a person responds to a flickering image of the cockroach. Link to original source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150667.

As Li and Zhang note, the adaptor shows "stable and continuous high levels of accuracy in both 'sender' and 'receiver' sides" (2016, p. 15)[2]. Accordingly, to address the rise of synergies and functionality, we focused on, first, the measures of cockroach sensitivity to micro-electronic prompts (cyborg response accuracy) and, second, human success in keeping the insect within boundaries (human success rate). Table 1 presents selected findings from those reported in detail in the original paper.

Although one cockroach reduces the human success rate, broadly, human "thought" sets off high cyborg response accuracy. Tongue in cheek, the authors mention cockroach three's "self-willingness" or, strictly, the role of extraneous variables. Crucially, given the human success of about 20%, the task is not easy. Given this fact[3], we treat variability as showing, first, the scope for learning and, second, marked individual differences. It is striking that human subject three has the most accurate EEG classification, the best cyborg responding, the highest success rate, and alone, some success with cockroach three. We infer that much depends on managing how the adaptor bridges between a human brain and the cockroach's antennae nerves (i.e., human-centered control of EEG input). In spite of cyborg tendencies, the cockroach is no automaton. In contrast, humans must learn to use the adaptor in task-specific ways. Since these require both motivation for success and a grasp of the problem (but *not* what to do), the techniques involve more than learning. Rather, one must ask how an adaptor shapes vicariant effects in a novel task.

TABLE 1 Success in controlling cockroach moves.

	Cyborg response accuracy (%)	Human success rate (%)
Cockroach 1	93.1	33.3
Cockroach 2	82.4	20.0
Cockroach 3	82.9	6.7

Even if training improves skills, techniques develop and, as the success rate shows, no knack emerges. While the device sends "instructions" to the cockroach (given high response accuracy), human "thought" is subtle. Far from being a means to an end or a functional tool, the engineered system empowers the subject as an adaptor. It brings the once impossible within reach as perceptual assessment becomes part of willing a cockroach to move. Given the device, a brain-in-the-system synthesizes the *ways* of adapting (see Figure 3). As in the classic work on Tetris, the engineered system prompts the self-fabrication of epistemic powers (Kirsh and Maglio, 1994). In spite of the device's novelty, the resulting techniques use "tacit and overt controlling capacities" that allow "purposeful pre-reflective (bio)mechanical execution" (Gahrn-Andersen and Prinz, 2021). Importantly, "willing a move" must *feel* like something (for the person-in-the-system). Hence, the pre-reflective can contribute to epistemic effects as a person with a brain-operating-in-a-wide system sets off tacit neuronal tinkering. In the terms of Gahrn-Andersen and Prinz (2021), the device affects a "state of being" through "subconscious adaptation and fine tuning of neuronal circuits" (p.

2 The develop a *control performance coefficient* to contrast system performance as compared to chance (or a control). They tested the mean CPC value of 0.616 ± 0.169 against the chance level (0.375) with a one tailed t test and found it was highly significant, citing a t $p < 0.0001$ ($t = 8.170$).

3 Given longitudinal data, we could not track the role of "watching and willing" or how "noise" affects classification of measures. We do not attempt that here.

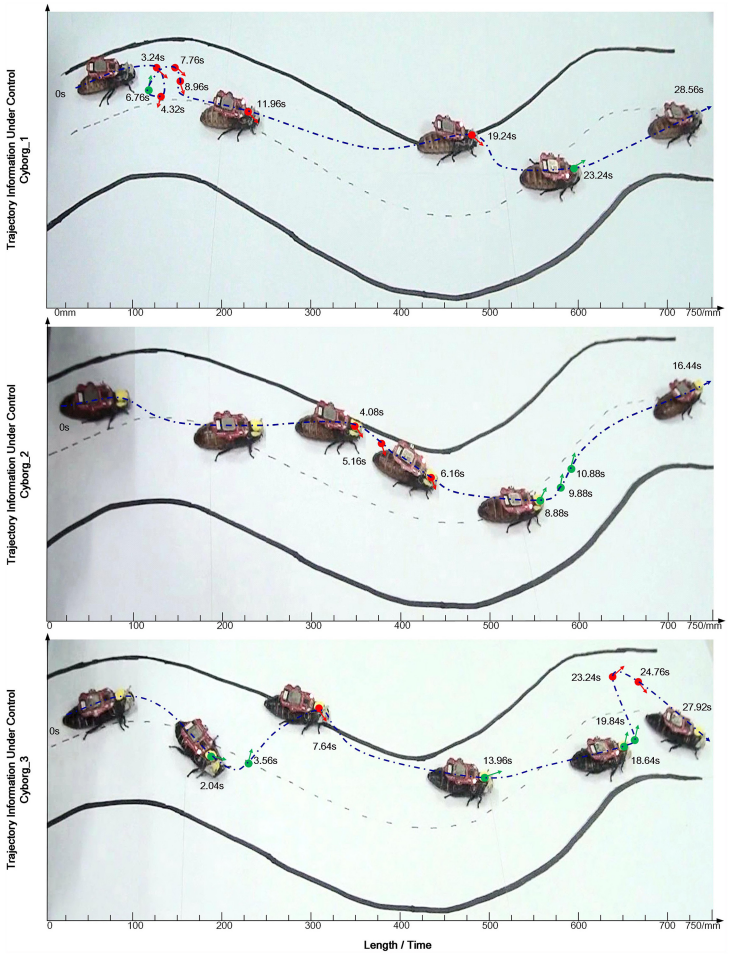

FIGURE 2
The trajectory of a cockroach moving on the S curve showing time taken. A green dot indicates a left-turn command; a red dot indicates a right-turn command. Link to original source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150667.

110). In short, reuse enables the brain to self-design techniques for human control of the cockroach[4].

While based on trial and error, the technique is not reducible to the "law of effect" (Dennett, 1975). Rather, as epistemic engineering, the brain gains functionality that acts as valid knowledge that is oriented to, not just a stimulus, but also the adaptor-person. Within the poly-centered system, the results attune the brain-in-the system to watching and willing. The cockroach "part" enables reinforcement to calibrate how a phenotype is extended by a system that couples an engineered adaptor, neural activity, and the pre-reflective. Hence, this constitutes natural evoneering. In the terms of Dennett's (2017) heuristics, the person needs more complexity than a Skinnerian agent but not the "inner environment" of its Popperian counterpart[5].

Rather, the brain reuses old tricks that link distributed agency with vicariance. Persons use wide systems such that, without knowing what they are doing, they bring purposefulness to learning. In Dennettian vein, one might call them Tolman agents who act with intent (i.e., *as if* they were purposeful)[6]. Just as in acting as a Morse operator (Cowley, 2019), the pre-reflective shapes techniques in a person part of a wide system. As in Tetris (Kirsh and Maglio, 1994), persons-cum-brains use the feel of *attending to the perceived*. Techniques use recursive trial and error to connect cognitive events with the feeling of what happens (Damasio, 1999). As a result of

4 Gahrn-Andersen and Prinz (2021) suggest that, for the human part of the system, the brain's enabling activity is part of the "pre-reflective." Since one feels about what one sees one needs no "representations". This is possible, they suggest, because hierarchies of molecular coding draw on (and, perhaps re-use) configurations of electromagnetic and cognitive patterns. The brain may combine the use of more meaningful peripheral elements with a computational core.

5 In Dennettian terms, Skinnerian agents link a history of reinforcement together with planning and selection such that, in some species, culturally transmitted replicators sustain "off-line" learning. Unlike Popperian agents, they lack "models" of the world: in developing an *ex novo* technique, one needs neither cultural replicators (instructions) nor a model that corresponds to an external environment. Presumably, the novel technique arises from a (coded) reconfiguring of neural sub-systems (or what Piaget calls accommodation) as well as reinforcement. Importantly, one need not know that one is controlling EEG input; change happens *for* a person (who can falsely believe they rely on "thought control").

6 The label is in Dennettian spirit. While alluding to Tolman (1932), we do not suggest that such agents act in accordance with his theory. Simply, they use the law of effect to act in ways that, seen from an intentional stance, appear purposeful.

actualizing practices, experimental subjects draw on brains to self-fabricate techniques that allow for reasonable task performance.

When the engineer adds vicariant systems (e.g., a screen and EEG device) to human-cockroach engagement, the human part of the system can direct "input" to the adaptor (refer to Figure 1). What is possible is transformed: natural evoneering enables a novel technique.

Over time, the subject's brain gives rise to techniques based on seeing how the cockroach moves. Far from reducing to stimulus-response or planned action, a living human subject uses "thoughts" as attending to how the seen sets off retrojecting. The anticipative results trigger learned parameters and EEG measures, which act as output *for* a cockroach. With training and experience, humans alter how the agency is distributed between the body, devices, and the cockroach. The human uses the pre-reflective – or: the conscious but not reflectively conscious – in the entirely innovative engagement with an engineered device. Given familiarity with a cockroach-in-the-system, the pre-reflective sets off prompts and thus vicariant effects. Cognizing is evoneered across a brain that attunes to a screen and EEG device as the person-adaptor gains know-how. As a result, pre-reflective experience triggers neurophysiological events or, loosely, "thoughts." In such a case, we meet the challenge set by Afeltowicz and Wachowski (2015): the emergence of intent (or the purposeful actions of the human) uses the interdependencies of a motivated poly-centric system. Novel behavior draws on a history that links the pre-reflective, neural activity, use of an adaptor, and contingent effects. The system's world-side resources (the adaptor-and-cyborg cockroach) use brain-side systems to shape the feeling of what happens to grant human subjects techniques. Hence, the case of minimal epistemic engineering relies on actualizing a social practice whose functionality appears to an outside observer (although the performer lacks any sense of how results are achieved).

4. Epistemic engineering in a working environment

Next, we turn to vicariant effects that arose when drones were introduced to a Danish utility company. Similar to the cyborg-cockroach approach, parts use epistemic engineering within a practical assemblage (Nail, 2017) that can be (partially) described by distributed cognitive systems[7]. The changes both draw on–and favor–vicariance as agents change both how they act and/or what they know. While natural evoneering occurs, in this case, agents often also gain a "grasp" of their place in changing public practices. As shown below, this applies especially to a system operator whose work is pivotal in the working environment. Drawing on the experience of other tasks (i.e., of a pre-drone task regime), he brings forth new possibilities. As a result, human participants grant systems and parts new functionality that, in practice, constitutes valid knowledge. They use an experience-based sense of events, or the feeling of what happens, to actualize practices. Furthermore, they discuss the results and use their talk to adjust later behavior, alter systemic function, and, thus, the use of parts, materials, and a task regime. In this case, there are no new intents. However, just as with the cockroach, the change reduces to neither planning nor the automatization of skills. Rather, it arises from *grasping* how systems can bring forth new kinds of functionality.

4.1. Pursuing vicariance in a Danish utility company

In Denmark, district heating supplies most urban environments and is used by 64% of all households. With such heating, hot water is pumped from combined heat and power plants through distributed stations to private homes, businesses, and public institutions. After reaching its destination (i.e., the radiators of the structure to be heated), the "used" water returns for re-heating *via* a network of pipes. While ideally closed, the system suffers from spillage and, for this reason, companies have to add make-up water (and consume extra energy). For this reason, to reduce, or prevent, such leakages without changing pipes, a crucial role falls to the work of the maintenance department. In 2016, the utility company in question began collaborating with a provider of drones that use thermographic cameras for leakage detection. The cameras readily detect the changes in heat radiation from water that is pumped at around 80°C: once the information is identified, heat radiation from underground pipes can be rendered "visible."

Many different practices[8] contribute to the maintenance of the pipe network. In this context, therefore, we stress that the introduction of drone technology has cascading consequences. Indeed, the prominence of leakage detection has vicariant effects across the company. To us, it appears that drone-based effects are transforming the mission of maintaining the pipe network. For now, we track innovation in a bundle of practices (i.e., maintaining the pipe network) that, in return, have fed both across other work and back into the use of drone-facilitated information in the maintenance department. In the subtask regime that has arisen, the use of drones (1) creates a novel task (i.e., thermographic leakage detection); and, (2) qualifies an existing one (e.g., the repairing of leakages) relates to the mission of maintaining the pipes. Unplanned changes thus have far-reaching consequences because existing work must both fulfill extant task regimes and, at once, alter in responding to use of drone cases. Hence, drones have become increasingly central to maintenance practice, changed relations between employees and external contractors, and prompted senior management to set a weekly target for dealing with drone cases. The vicariant effects are unplanned because, rather than integrate the drone task regime with extant practices, they have had to be improvised. They have been brought in piecemeal both to supplement general operations (i.e., "non-pipe related maintenance tasks" such as the change of manhole covers) and in changing the pipe network maintenance (e.g., the repairing of alarm threads in certain pipe types). For ease of exposition, we now draw a comparison with the minimal case by identifying the outward flow of vicariant effects.

Over time, seeing the images triggers a cascade of vicariant effects (leading to both intra-organizational change and effects on sub-contractor operations) (see Figure 4). Under the old task regime,

7 *Assemblage* is used in translations of Deleuze and Guattari who apply the term to characterize parts that co-function neither in ways predetermined to fit an already-conceived design nor a random collection of things see, Nail (2017). Where parts align with functions they can be described as a distributed cognitive system.

8 Indeed, the utility company's history of proving district heating goes back to 1925.

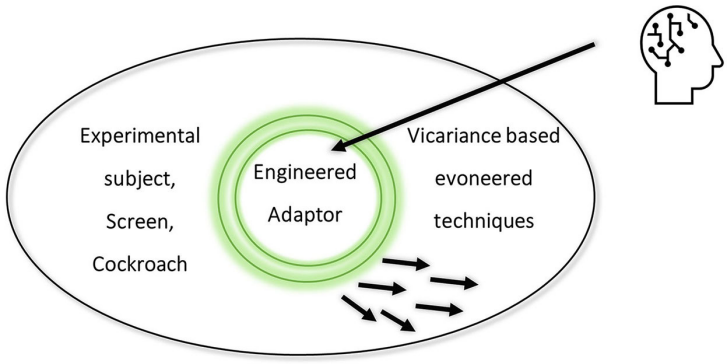

FIGURE 3
The experimenter (human head) designs a system with an engineered adaptor. As a whole adaptor system, the person, screen, and cockroach co-constitutively draw on natural evoneering.

FIGURE 4
Screenshot of Teraplan.

decisions about repairing leaks drew largely on contingencies. Since the utility company had no means of seeking out leakages, they relied on when, for instance, a vigilant citizen found green water in their basement (the make-up water has added green color) or if, following a snowfall, an expert noted melted snow above an underground heating source. Hence, drones brought a new order to their work[9]. Furthermore, since they have proved both reliable and efficient, the leakages could have potentially overwhelmed the department's financial, human, and other resources. As one senior manager says: "The drones give us knowledge of leakages that it would otherwise take 10–15 years to gather" (Senior manager). As so often with digital solutions, the accumulation of data demands epistemic engineering and, at once, sets off epistemic change. Having seen that drones bring about new functionality, senior management set the target of addressing 5 new drone cases each week.

4.2. Drone task regime: Screening and managing of incoming data

The utility company uses a drone service provider as a semi-autonomous assemblage that provides images based on the specialized software (see Figure 4). Given a technical specification, the parts couple tightly with the company's task regime: employees quickly established the routines based on the classification of

9 Here, we are looking beyond leakages that are automatically reported by the alarm threads in certain pipe types. The drones have been introduced with the purpose of spotting leakages in pipes that do not come equipped with such threads.

suspected leakages. The service provider package includes (a) aerial surveillance of areas of the city and then (b) thermographic images from the surveillance operations supplied to through licensed, custom-built software: *Teraplan*. In the case of (b), the Teraplan data are the drone provider's extension of Google Maps to classify the suspected leakages on a certainty scale (*viz*. As are most certain, Bs less so; and Cs are call for further examination). Furthermore, the user can turn software layers on and off (i.e., to focus on the thermographic layer, Google Maps satellite photographs or the utility company's network of pipes; refer to Gahrn-Andersen, 2020). Plotting of the suspected leakages is performed manually by the drone operator who screens thermographic images while using a depiction of the utility company's network of pipes.

For the maintenance department, Teraplan sets off vicariant effects. Since these must be monitored and managed, the program is shaping an unplanned task regime. In this context, the role of the system operator takes on new importance. Above all, this is because the role now combines extant knowledge and skills (e.g., knowledge of the streets of the city) with a grasp of what Teraplan shows. Drone-based information combines with personal knowledge that draws on the company's own Geographical Information System (GIS). Rather as with the cyborg cockroach, images-cum-software demand that the system officers attune to the output of Teraplan. Bodies function as parts of an adaptor (just as, elsewhere, a Morse operator's body comes to act as an adaptor, see Cowley, 2019). While we later highlight contrasts, parties close to the software are required to develop techniques (not described here) that, oddly, bring new understanding to the old experiences. The resulting decision-making alters the parameters of action and, thus, company practices. We begin with how, given the accuracy of leakage detection, the system operator sets off epistemic engineering. Given his grasp of how drone-based information bears on the wide system, he has to (1) verify the leakage indicated and (2) initiate repairing by forwarding relevant information to the sub-contractor.

Since Teraplan indications of leakages are accurate, the system operators developed a distinctive routine. They link the output to professional knowledge and the utility company's GIS system to set off vicariant effects across the whole system (i.e., the rest of the maintenance department, relevant contractors, the municipality, and private citizens). The resulting epistemic engineering is achieved by acting in ways that favor leakage repair: just as with the cockroach, epistemic change arises as parts of the assemblage exert co-control. These are funneled by how the service provider's coders process raw data and, above all, the system operator's validations and decisions. In what follows, we focus on suspected leakages that are classified as As. While the classification has identified hundreds of successful cases, there are also errors. For example, one A identified ground that had been heated up by a parked bus, and in another case, it showed clamping close to the surface as shown on the utility company's GIS depiction of pipes. Accordingly, the system operator makes an experience-based assessment of each leakage: information from Teraplan is verified by a double check or, as a system operator says: "[The drone] doesn't know what is underground. The GIS [Geographical Information System] does." While Teraplan can show whether a suspected leakage is close to a pipe, the GIS system adds detailed information about each pipe's type, dimensions, exact lengths, etc. Hence, the system operators compare the Teraplan images with the information from the GIS. They use personal knowledge to identify false positives such as when increased thermographic radiation on clampings does *not* show a leaking pipe. Hence, one system operator, a smith with years of hands-on experience, stresses the need for fine comparisons between images from the two information systems:

> As long as we have these two systems [i.e., Teraplan and GIS] like this, it is fairly simple to work with them. Because I also think that we need to keep ourselves from accessing this one [i.e., the GIS] too much. In spite of it, it is a webpage which runs constantly, and our GIS system is so massively huge, you know. It is a way heavier system [than the drone operator's software]

Having double-checked the Teraplan data with the GIS, the system operator also draws on his own experience in deciding when to authorize the utility company's contractors to start on any given case. As confirmation, the contractor begins with a preliminary digging to validate the accuracy of the spot identified. Additional measures require that a system operator or contractor visits each suspected leak and verifies the results using a handheld thermographic camera. However, given the precision of coding As, this procedure has become little more than a formality. Leaving aside work with Bs (let alone Cs), we now turn to how, in the second part of the drone task regime, important contrasts arise with the cockroach case. This is because, as vicariant effects fan out from the system operators, they lose predictability: managing repairs requires entangled links between organizational settings and, thus, care in adapting parts of the assemblage as one manages distributed agency.

4.3. A secondary dimension of the assemblage: How the repairs are managed

Whereas opening the drone case has become part of a routine, the subsequent management of repairs is rather loosely structured. Much depends on a weekly "damage meeting" [Da. Havarimøde] where the maintenance work is organized. The meeting enables drone task work while also dealing with both pipe and non-pipe-related maintenance. Each case is given status updates and, where works are not progressing, solutions are brought forward. The logic of each repair is roughly this: (1) the contractor applies to the municipality for permission to dig; (2) affected customers are notified of heating disruption; (3) once a leakage is dug free, its extension is approved by a system operator (who might also chose to temporality close the hole). Later, when the pipe can be replaced by contracted pipe specialists, (4) the utility company sends out a technician to turn of the water. In step (5), the contractor replaces a section of the pipe, and, in (6), the utility company technician restores the flow. Next, in (7), the digging team fills up the hole, lays new asphalt, and removes barriers and signs. Finally, in (8), the utility company technician fills out a "damage report" [Da. Havarirapport] that documents the works and serves to update information in the GIS. In actual circumstances, of course, the progression can be negatively affected by the factors such as staff shortage, an overload of cases, or unforeseen events (e.g., frost that makes digging difficult). In what follows, we

present two drone cases reported at a damage meeting held on 26 March 2019:

Drone case 1:

> Digging commences in week 49. 11-12-2018: Digging in week 50 because we did not manage in week 49. 18-12-2018: Digging commences 19.12.18. 08-01-19: The digging permit [which is temporary and issued by the municipality] has been reevoked due to expiration. A new hearing phase has started. 15-01-2019: hearing is ongoing 22-01-2019: hearing ongoing. 29-01-2019: Digging permit received, commencing in week 6.05-02-2019: Digging d.6/11. 12-02-2019: Digging. 19-02-2019: Waiting due to parked car. 26-02-2019: Still waiting because of the car. 05-03-2019: Waiting due to parked car. 12-03-2019: digging completes in this week 11. Is being planned. 19-03-2019: digging finished. 26-03-19 status unknown.

Drone case 2:

> Ready for [the contractor]. Contact the customer prior to commencing. 29-01-2019: Shooting pipe [a type of pipe] 22.05-02-2019: Expected beginning in week 7. 12-02-2019: [Manager 2] follows up with [digging contractor] in relation to the commencing. 19-02-2019: Commencing Friday 22/2. 26-02-2019: commencing week 9. 05-03-2019: Commencing 06.03.19 12-03-2019: A Greek in place [a term for a temporary repair of the leakage] 12/3. Expected clearance digging week 11. 19-03-2019: [Utility company technician] is to contact [digging contractor] regarding eventual repositioning of the plug. On the agenda for the supervision meeting 22/3. 26-03-2019 digging continues

Notably, the meeting focused on 18 drone cases: as was now clear, the utility company had inadvertently caused a bottleneck. This is because, without having any means of tracking vicariant effects, senior management had introduced a target of five drone cases a week. Given the unplanned nature of the process, additional drone cases were issued to contractors on 12 March and, by the time of the meeting, the bottleneck had been developing for a month. Indeed, for reasons that cannot be discussed here, the continuous addition of new drone cases led to unexpected difficulties for, above all, the digging contractor. Subsequently, the utility company was to react by temporally suspending its "five leakages per week" policy. The two cases serve to illustrate the problems and give a sense of what, precisely, is meant by saying that drones led to epistemic engineering as systems and parts developed functionality that, for those in the company, constitute valid new knowledge.

In the first case, 3 months had passed in progressing from steps 1–2 to the operational repair procedure. This was due to two unforeseen tasks: (a) renewal of the digging permit and (b) the need to remove a parked car which, in fact, led to a 2-month delay before digging could begin (the reason for this was that the company then faced issues with expired digging permits and material and manpower shortage). Whereas, the need to reapply for the permit is a dysfunctional element due to shoddiness and lack of manpower, the second case is a common contingency that, in this case, led to a serious delay. By placing a "Greek" on the pipe, the utility company successfully completed step (3). Yet, since more coordination (i.e., a "supervision meeting") was needed, an emergency *ad hoc* meeting was called to deal with cases that were piling up because a contractor had fallen far behind schedule. In this particular case, both the contractor and the utility company had overestimated the duration of repairs, and conversely, underestimated how maintenance operations would be influenced by environmental factors.

Unlike the minimal system, the utility company's systems are, at once, organized and deeply entangled. They arise in a poly-centered unit that includes people with very variable understanding. The results have the indeterminacy of *systemic assemblages* (Gahrn-Andersen, 2020) that are: (1) open to social, market, and technological change; (2) enable drones and information to produce functionality; and (3) bind the causal, the biological and social. As we see, drone functionality is fully entangled within organized life: it includes, first, coders (and drone operators) who plot useful data in Teraplan; second, it has made the system operator who uses the software into an "adaptor" like a person with a verifying/facilitating role. However, the assemblage must cope with not only drone-derived data but also seemingly drone-independent repercussions that are conceptualized around the tasks of repair. Indeed, given poly-centered control, as in similar organizations, the utility company uses a hierarchical structure to maintain institutional control (e.g., through damage meetings). In clarifying how parties manage epistemic engineering, therefore, we draw contrasts to the minimal system. Whereas, the human-cockroach adaptor is encapsulated, Teraplan makes the system operator into an adaptor whose functionality disseminates. To an extent, diverse, loosely coupled systems demand from the other human parts that they adjust their ways of acting (and develop novel techniques). Above all, skillful agents (the drone operator's coders and the utility company's system operators) determine the company's function and operation. Hence, in moving from a drone-specific task regime to the maintenance task, the task coupling becomes looser and, at times, decouples (at least in part). In such cases, additional supervision meetings are needed (cf. Drone case 2). In bringing order to such events, we now consider the implications of recognizing how adaptors set off vicariant effects. We stress that, since epistemic change is incorporated into action, talk, and routines, human cognition can use how intents and epistemic change arise in socially organized wide systems (refer to Figure 5).

5. Organized humans: A systemic view

Complex systems such as toy locomotives and galaxies contrast with the bodies that subserve human knowing. As Bateson (1979) notes, "the toy locomotive may become a part in the mental system which includes the child that plays with it, and the galaxy may become part of the mental system which includes the astronomer and the telescope (1979, p. 104)." In his terms, objects are not thinking subsystems in larger minds but, rather, nature evolves as observers (or knowers) use *relationship*s. Overlooking entropy reduction, he suggests that these arise "between two parts or between a part at time 1 and the same part at time 2 (p. 106)" and activate a third component such as a sensory end organ. The receiver "responds to is a *difference* or a change" (Bateson, 1979). Receipt of the differences makes a difference for a system. In parallel, for Giere (2011), there is an asymmetry of knowing

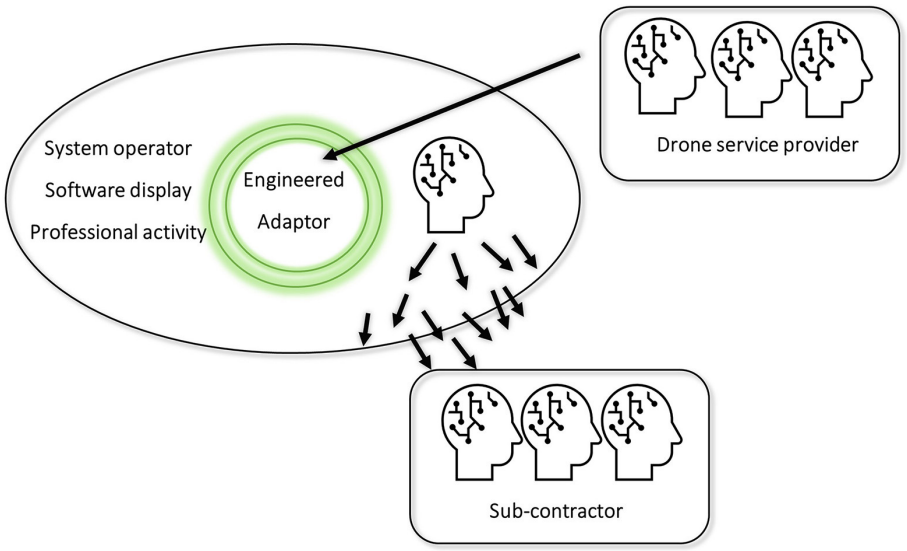

FIGURE 5
The drone service provider (outer source) enables a system with an engineered adaptor. As a whole adaptor system, the system operator, the software display, and professional activity set off epistemic change.

and cognizing. As illustrated by the Hubble telescope, whereas cognitive outputs (e.g., images from space) derive from the whole system, only human parts can *know* anything. This asymmetry is fundamental because of the clear implication that bodies use cognitive input to create an epistemic output (differences that makes a difference for a system and/or its parts). In Bateson's terms, distributed systems use "differences" or information that the doings of living parts transform into knowledge and know-how (as things happen). Yet, Giere leaves aside *how* "receiving" can prompt coming to know. In addressing this in humans, we suggest that knowledge arises in wide systems as living parts reduce entropy, simplexify (Cowley and Gahrn-Andersen, 2022) and make use of adaptor systems.

As epistemic actors, humans both receive and process information (or perceive differences) as they exert control over the results. In focusing on how cognition binds human understanding with the deliverances of wide systems, we take a systemic view (Cowley and Vallée-Tourangeau, 2013, 2017). As with the cockroach controller or the drone system operator, epistemic change uses systemic interdependency. Whereas, cognizing pertains to a whole system, *knowing* concerns Giere's (2004) "human cognitive agent" or, simply, a living human being. The move resolves the collective-individual tension noted by Baber (2010), Perry (2013), and Jones (2013) by making artifacts and language part of a distributed agency. As shown by Fioratou and Cowley (2009), for example, insight problems are solved as bodies are nudged to abstract "aspects" from lived experience. In Cowley and Vallée-Tourangeau's (2017) terms, primates "notice things" by drawing on what is called the principle of cognitive separability (PCS). In noticing, we take distance from body-world engagement as doings attune to *aspects* of things. In tool use, for example, we "try" things out and, with experience, learn from practice (Donald, 1991). Given distancing (and the PCS), a contingency can prompt *seeing* a solution (Ball and Litchfield, 2017) or problem-solving can be triggered by the aesthetics of symmetry (Steffensen et al., 2016). Positing the PCS both clarifies epistemic outcomes and also shows the cognitive value of attending to emplaced experience. Together with distancing, one can generate intent and epistemic change using interactivity (Kirsh, 1997; Gahrn-Andersen, 2019), resonating with pico-dynamics (Blair and Cowley, 2003) or striving for cognitive events (Steffensen, 2013). The PCS links routine performance with higher cognitive functions (Cowley and Vallée-Tourangeau, 2017). Yet, appeal to a principle leaves aside how living parts of wider systems change parameters with epistemic effect. After all, only *some* events shape techniques and only expertise can derive useful outcomes from systemic interdependencies. It follows that distributed systems do not just self-sustain but, just as importantly, co-function as persons, brains, and bodies generate epistemic change. Given distancing, attention, and emplacement, people draw on a life history to exhibit powers associated with what Madsen (2017) calls multi-scalar temporal cognition. In a Mafia setting, for example, a mother may desecrate her child's "informer's grave" (Neumann and Cowley, 2016). Coming to "know" the appropriacy of such action eludes both neurophysiological or convention-based accounts (i.e., micro- or macro-explanation). Rather, the desecration attests to an organized domain where human agents link the micro with the macro. As a member of the Mafia world, the mother is concerned with neither a task nor a distributed cognitive system. Damaging her child's grave is inexplicable by accounts based on either interaction history or normative social roles. Rather, events presuppose a public space of action where wider systems operate as constraints on neurophysiology and, thus, action: adjustments unite public appearances (and responding to them) with the macro-social and the bio-behavioral. Formally, one can posit the three co-functioning dimensions (Secchi and Cowley, 2016, 2021; Secchi et al., 2023) known as the Ms (macro, micro, and meso). In peer review, for example, a reviewer drives epistemic change by drawing on organized structures, individual prompts, and judgments of what is likely to be perceived as having scientific value (Secchi

and Cowley, 2018). Tasks and cognitive ecosystems become part of a meso-domain–a public space of unending, structural change.

A focus on structural change privileges systemic interdependency. As in the Mafia case, behavior is irreducible to interaction. People simplexify or reduce entropy by drawing on retroactive processes. They amalgamate past experiences with a lived now both in willing cockroach movement (using techniques) and binding Teraplan images with "knowing" the streets shown by the GIS software. Within a meso-domain, one acts as *a person in the system* (Fester-Seeger, 2021). As parts of wide and distributed systems, in Bateson's (1979) terms, people recognize the differences and enact news. As Hutchins sees, they reduce entropy and, we add, set off vicariant effects that *make* differences. The claim matters in that it addresses Afeltowicz and Wachowski's (2015) objection to the distributed perspective. Intents can be public, multiscalar effects that embody epistemic changes. In the cockroach experiment, an engineered adaptor prompts an experimental subject to develop purposeful behavior. While brain-enabled, *contra* Afeltowicz and Wachowski (2015), thoughts need, not a neural mechanism, but a special *way* of "looking while willing." The brain creates novel structures (techniques) within a wide system where a person becomes part of an adaptor system that controls the brain-cockroach whole. In the utility company, a system operator achieves epistemic outcomes by retrojecting the experience of terrain onto a software display. As an expert, he can see that Teraplan shows a bus stop that is "too far" from the side of the road. In such a case, expertise can prompt one to challenge evidence. Cognizing thus arises in the meso-domain of an extended system: this is where the experimental subject makes the cockroach turn and the system operator decides to check an intuition at the site specified. While brain-enabled, the action is reliant on public cues; the brain's role is, not to control, but to grant a sense of purpose (i.e., as in a Tolman agent). In the wide system, the cockroach controller amalgamates changing impressions (the system in the person) with increasingly effective action (independent of belief). In parallel, organized routine co-functions with equipment to form a system operator's intuition. Furthermore, while the PCS plays no role in the action, the techniques presuppose an enlanguaged world (refer to Cowley and Gahrn-Andersen, 2022) where actions make sense: this enables a person in the system to see what can be done or grasp what one is meant to do.

Sensitivity to the moment is the hallmark of social organizing. It allows the persons to attribute a public (or "relevant") sense to events and, thus, establish vicariant effects. Hence, living systems use systemic interdependencies to shape the "outward spread" of knowing. In the drone case, the spread affects a range of stakeholders as persons reduce entropy through epistemic engineering. While using routines and cultural ecosystems, parties also develop techniques and act to simplexify. Without knowing what they are doing (or explicit training), they alter both whole system functions and also those of bodies and living persons. Epistemic change can reveal what one "should" do or prompt a grasp of the possible. Often, experience, expertise, and techniques bind with what linguists call entrenchment (Cowley, 2017; Schmid, 2020). The resulting judgments use, not a faculty of reason, but how practical know-how unfolds in an enlanguaged world. Experienced individuals gain capacities for reliable judgments and making use of docility (Secchi, 2016). In the utility company, these qualities–not just routine use of systems–ensured a smooth transition to drone use in pipe maintenance. A well-organized systemic whole ensures that drone-based information is currently driving the reorganization of maintenance work (Gahrn-Andersen, 2020). As change spreads, people link bodily feel and expertise to causal systems that set off cumulative practical effects. The equipment serves, not just directly, but also to improvise new material and institutional relations (i.e., by setting boundary conditions on sensitivity to linguistic semiotic resources). The vicariant effects enable the teams and individuals to (a) self-empower; (b) reorganize; (c) influence each other; and (d) alter routines. Parties gain expertise, skills, and ways of drawing on the system. Thus, while many new issues arise (e.g., reorganizing supply and budgeting needs), the drone study also shows how resilient organizations and individuals gain from cascades of epistemic change.

6. Epistemic engineering

Emphasis on systemic interdependencies plays down the role of organism-centered control. Indeed, the radical potential of the systemic view lies in bringing a constructive role to distributed systems. As we have argued, they enable humans to generate intents, epistemic effects, and collective knowing: often persons *make* differences using wide systems to set off vicariant effects. During routines or practices we enact and mimic adaptor systems that trigger epistemic change. Hence, agency and tasks are reciprocally related. The view clarifies how wide systems contribute to social intelligence in lemurs (Jolly, 1966; Sterelny, 2007), navigating a ship (Hutchins, 1996), or using "thought control" over a cyborg cockroach. In hominins, neural plasticity co-evolved with new variation in cognitive performances: at times, we attend closely and, at others, we distance ourselves and, given hints, gain insights ("perhaps a bus warmed the ground"). In part, this is due, we suggest, to the principle of cognitive separability that allows us to notice potential value in the contingent. Indeed, without it, there would be no flexible-adaptive tool use or amalgamation of social regularities and irregularities. By implication, the epistemic novelty of hominins may derive from our use of distributed agency. With the rise of artifice, humans come to draw on, not just bodies, but also reciprocal relations within wide systems and across practices.

In an enlanguaged world, vicariant effects contribute to intents, routines, and practices. In the "minimal" case, a person-in-the-system sets off epistemic change by purposefully moving a cockroach. In the system, looking-and-willing reduces entropy as a brain adapts to the engineered adaptor. In the utility company, epistemic change reaches beyond techniques as drones cum Teraplan software enable a system operator to set off a cascade of effects. In this case, epistemic engineering prompts people to see opportunities and, over time, figure out what to do: while requiring neural re-use and control, the power of self-sustaining systems (and the meso-domain) lies in generating useful knowledge. By enabling adaptor systems, we use epistemic effects to get things right. Without any foresight, people link entropy and the pre-reflective with the hints and nudges of an enlanguaged world. Within interdependent and distributed systems, vicariant effects enable epistemic change, self-empowerment, new uses of equipment, co-creativity, and variation in routines. We, therefore, submit that much is gained from teasing apart living agency from that pertaining to supra-systems, tasks, and routines. The radical move allows cognitive powers to use, not only bodies, brains, and organism-environment coupling, but how human life cycles

serve in making differences. The biosocial resources of wide systems can be used to ensure that distributed control sets off vicariant effects whose parameters function to *construct* epistemic change. In short, while selection filters novelty, non-linear change transforms the knowable. By hypothesis, then, epistemic engineering is an evolutionary principle that may well apply across the living world.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

Both authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

Both authors acknowledge the funding awarded by the Velux Foundation of the project Determinants of Resilience in Organizational Networks (DRONe) (grant number 38917). Moreover, we are grateful to Sune Nielsen and Bo Jensen Møller for permitting us to use the Teraplan screenshot. Thanks also to Li Guangye for feedback on early thinking and Marie-Theres Fester-Seeger for insightful comments on a draft.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Afeltowicz, Ł., and Wachowski, W. (2015). How far we can go without looking under the skin. The bounds of cognitive science. *Stud. Log. Gramm. Rhetoric* 40, 91–109. doi: 10.1515/slgr-2015-0005

Anderson, M. L. (2010). Neural reuse: a fundamental organizational principle of the brain. *Behav. Brain Sci.* 33, 245–266. doi: 10.1017/S0140525X10000853

Baber, C. (2010). Distributed cognition at the crime scene. *AI Society* 25, 423–432. doi: 10.1007/s00146-010-0274-6

Baber, C., Parekh, M., and Cengiz, T. G. (2014). Tool use as distributed cognition: how tools help, hinder and define manual skill. *Front. Psychol.* 5, 116. doi: 10.3389/fpsyg.2014.00116

Ball, L. J., and Litchfield, D. (2017). "Interactivity and embodied cues in problem solving, learning and insight: further contributions to a "theory of hints," in *Cognition Beyond the Brain*, eds S. J. Cowley and F. Vallée-Tourangeau (Springer, London), 115–132.

Bateson, G. (1979). *Mind and Nature: A Necessary Unity*. New York, NY: Dutton.

Berthoz, A., and Tramus, M. (2015). Towards creative vicariance: interview with Alain Berthoz. *Hybrid.* 2, 1–7. doi: 10.4000/hybrid.1325

Blair, G., and Cowley, S. J. (2003). Language in iterating activity: microcognition re-membered. *Alternation* 10, 132–162.

Cheon, H. (2014). Distributed cognition in scientific contexts. *J. Gen. Philos. Sci.* 45, 23–33. doi: 10.1007/s10838-013-9226-4

Cowley, S. J. (2011). *Distributed Language*. Amsterdam: John Benjamins.

Cowley, S. J. (2014). Linguistic embodiment and verbal constraints: human cognition and the scales of time. *Front. Psychol.* 5, 1085. doi: 10.3389/fpsyg.2014.01085

Cowley, S. J. (2017). "Entrenchment: a view from radical embodied cognitive science," in *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*, ed H. J. Schmid (Berlin: Walter de Gruyter), 409–434.

Cowley, S. J. (2019). Wide coding: tetris, morse and, perhaps, language. *BioSystems* 185, 104025. doi: 10.1016/j.biosystems.2019.104025

Cowley, S. J., and Gahrn-Andersen, R. (2022). Simplexifying: harnessing the power of enlanguaged cognition. *Chin. Semiot. Stud.* 18, 97–119. doi: 10.1515/css-2021-2049

Cowley, S. J., and Vallée-Tourangeau, F. (2013). "Systemic cognition: human artifice in life and language," in *Cognition Beyond the Brain* (London: Springer), 255–273.

Cowley, S. J., and Vallée-Tourangeau, F. (2017). "Thinking, values and meaning in changing cognitive ecologies," in *Cognition Beyond the Brain*, *2nd Edn* (London: Springer), 1–17.

Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York, NY: Harcourt Brace and Company.

Dennett, D. C. (1975). Why the law of effect will not go away. *J. Theory Soc. Behav.* 5, 169–187. doi: 10.1111/j.1468-5914.1975.tb00350.x

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. New York, NY: WW Norton and Company.

Donald, M. (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge MA: Harvard University Press.

Edelman, B. J., Meng, J., Suma, D., Zurn, C., Nagarajan, E., Baxter Cline, C., et al. (2019). Noninvasive neuroimaging enhances continuous neural tracking for robotic device control. *Sci. Robotics* 31, eaaw6844. doi: 10.1126/scirobotics.aaw6844

Enfield, N. J. (2013). *Relationship Thinking: Agency, Enchrony, and Human Sociality*. Oxford: Oxford University Press. Enfield, N. J., and Kockelman, P. (2017). *Distributed Agency*. Oxford: Oxford University Press.

Fester-Seeger (2021). *Presencing: Rhythm and Human Cognitive Agency* (Unpublished PhD dissertation) University of Southern Denmark.

Fioratou, E., Chatzimichailidou, M. M., Grant, S., Glavin, R., Flin, R., and Trotter, C. (2016). Beyond monitors: distributed situation awareness in anaesthesia management. *Theor. Issues Ergon. Sci.* 17, 104–124.

Fioratou, E., and Cowley, S. J. (2009). Insightful thinking: cognitive dynamics and material artifacts. *Pragmat. Cogn.* 17, 549–572. doi: 10.1075/pc.17.3.04fio

Fodor, J. A. (1975). *The Language of Thought*. Cambridge MA: Harvard University Press.

Froese, T., and Ortiz-Garin, G. U. (2020). Where is the action in perception? An exploratory study with a haptic sensory substitution device. *Front. Psychol.* 11, 809. doi: 10.3389/fpsyg.2020.00809

Gahrn-Andersen, R. (2019). Interactivity and languaging: how humans use existential meaning. *Chin Semiotic Stud.* 15, 653–674. doi: 10.1515/css-2019-0033

Gahrn-Andersen, R. (2020). Making the hidden visible: handy unhandiness and the sensorium of leakage-detecting drones. *Senses Soc.* 15, 272–285. doi: 10.1080/17458927.2020.1814563

Gahrn-Andersen, R. (2021). Conceptualizing change in organizational cognition. *Int. J. Organ. Theory Behav.* 24, 213–228. doi: 10.1108/IJOTB-07-2020-0122

Gahrn-Andersen, R., and Prinz, R. (2021). How cyborgs transcend Maturana's concept of languaging: A (bio)engineering perspective on information processing and embodied cognition. *Rivista Italiana di Filosofia del Linguaggio* 15, 104–120. doi: 10.4396/2021204

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception: Classic edition*. Boston, MA: Psychology Press.

Giere, R. N. (2004). The problem of agency in scientific distributed cognitive systems. *J. Cogn. Cult.* 4, 759–774. doi: 10.1163/1568537042484887

Giere, R. N. (2011). Distributed cognition as human centered although not human bound: reply to Vaesen. *Soc. Epistemol.* 25, 393–399. doi: 10.1080/02691728.2011.605550

Humphrey, N. K. (1976). "The social function of intellect," in *Growing Points in Ethology* (Cambridge: Cambridge University Press), 303–317.

Hutchins, E. (1996). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Hutchins, E. (2001). Distributed cognition. *Int. Encyclo. Social Behav. Sci.* 2068–2072. doi: 10.1016/B0-08-043076-7/01636-3

Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philos. Psychol.* 27, 34–49. doi: 10.1080/09515089.2013.830548

Hutchins, E. (2020). "The distributed cognition perspective on human interaction," in *Roots of Human Sociality*, eds S. Levinson and N. Enfield (London: Routledge), 375–398.

Hutto, D. D., Kirchhoff, M. D., and Myin, E. (2014). Extensive enactivism: why keep it all in? *Front. Hum. Neurosci.* 8, 706. doi: 10.3389/fnhum.2014.00706

Jolly, A. (1966). Lemur social behavior and primate intelligence: the step from prosimian to monkey intelligence probably took place in a social context. *Science* 153, 501–506. doi: 10.1126/science.153.3735.501

Jones, P. (2013). "You want a piece of me? Paying your dues and getting your due in a distributed world," *in Cognition Beyond the Brain: Interactivity, Computation and Human Artifice*, eds S. J. Cowley and F. Vallée-Tourangeau (Dordrecht: Springer).

Kirsh, D. (1997). Interactivity and multimedia interfaces. *Inst. Sci.* 25 —96. doi: 10.1023/A:1002915430871

Kirsh, D., and Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cog. Sci.* 18, 513–549. doi: 10.1207/s15516709cog1804_1

Latif, T., and Bozkurt, A. (2012). "Line following terrestrial insect biobots," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (San Diego, CA: IEEE), 972–975.

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Li, G., and Zhang, D. (2016), ≪Brain-computer interface controlled cyborg: establishing a functional information transfer pathway from human brain to cockroach brain≫. *PLoS ONE* 11, e0150667. doi: 10.1371/journal.pone.0150667

Lintern, G. (2007). "What is a cognitive system?," in *2007 International Symposium on Aviation Psychology* (Dayton, OH).

Madsen, J. K. (2017). "Time during time: multi-scalar temporal cognition," in *Cognition Beyond the Brain* (London: Springer), 155–174.

Malafouris, L. (2013). *How Things Shape the Mind*. Cambridge, MA: MIT Press.

Malafouris, L. (2019). Mind and material engagement. *Phenomenol. Cogn. Sci.* 18, 1–17. doi: 10.1007/s11097-018-9606-7

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman.

Michaelian, K., and Sutton, J. (2013). Distributed cognition and memory research: history and current directions. *Rev. Philos. Psychol.* 4, 1–24. doi: 10.1007/s13164-013-0131-x

Nail, T. (2017). What is an assemblage? *SubStance* 46, 21–37. doi: 10.1353/sub.2017.0001

Neumann, M., and Cowley, S. J. (2016). "Modeling social agency using diachronic cognition: learning from the Mafia," in *Agent-Based Simulation of Organizational Behavior: New Frontiers of Social Science Research*, eds D. Secchi and M. Neumann (Springer, London), 289–310.

Pennycook, A. (2017). *Posthumanist Applied Linguistics*. London: Routledge.

Perry, M. (2013). "Socially distributed cognition in loosely coupled systems," in *Cognition Beyond the Brain*, eds S. J. Cowley and F. Vallée-Tourangeau (Springer, London), 147–169.

Rogers, Y. (1997). *A Brief Introduction to Distributed Cognition*. Available online at: http://www.id-book.com/fourthedition/downloads/chapter%208%20dcog-brief-intro.pdf (accessed June 1, 2022).

Ross, D. (2012). "Coordination and the foundations of social intelligence," in *The Oxford Handbook of Philosophy of Social Science* (Oxford University Press), 481.

Schmid, H. J. (2020). *The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.

Searle, J. (1992). *The Rediscovery of Mind*. Cambridge, MA: MIT Press.

Secchi, D. (2016). "Boundary conditions for the emergence of "Docility" in organizations: agent-based model and simulation," in *Agent-Based Simulation of Organizational Behavior*, eds D. Secchi and M. Neumann (Cham: Springer). doi: 10.1007/978-3-319-18153-0_9

Secchi, D., and Cowley, S. (2016). "Organizational cognition: what it is and how it works," in *16th Annual Conference of the European Academy of Management: Manageable Cooperation?*. Paris: European Academy of Management.

Secchi, D., and Cowley, S. J. (2018). Modeling organizational cognition: the case of impact factor. *J. Artif. Soci. Soc. Simul.* 21, 1–13. doi: 10.18564/jasss.3628

Secchi, D., and Cowley, S. J. (2021). Cognition in organizations: what it is and how it works. *Eur. Manag. Rev.* 18, 79–92. doi: 10.1111/emre.12442

Secchi, D., Gahrn-Andersen, R., and Cowley, S. J. (2023). *Organizational Cognition: The Theory of Social Organizing*. London, New York, NY: Routledge.

Steffensen, S. V. (2013). "Human interactivity: problem-solving, solution-probing and verbal patterns in the wild," in *Cognition Beyond the Brain*, eds S. Cowley and F. Vallée-Tourangeau (London: Springer). doi: 10.1007/978-1-4471-5125-8_11

Steffensen, S. V., Vallée-Tourangeau, F., and Vallée-Tourangeau, G. (2016). Cognitive events in a problem-solving task: a qualitative method for investigating interactivity in the 17 Animals problem. *J. Cog. Psychol.* 28, 79–105. doi: 10.1080/20445911.2015.1095193

Sterelny, K. (2007). Social intelligence, human intelligence and niche construction. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 362, 719–730. doi: 10.1098/rstb.2006.2006

Tolman, E. C. (1932). *Purposive Behavior in Animals and Men*. New York, NY: Appleman-Century.

Turvey, M. (1992). Affordances and prospective control: an outline of the ontology. *Ecol. Psychol.* 4, 173–187. doi: 10.1207/s15326969eco0403_3

Vallée-Tourangeau, F., and Wrightman, M. (2010). Interactive skills and individual differences in a word production task. *AI Soc.* 25, 433–439. doi: 10.1007/s00146-010-0270-x

Weibel, N., Fouse, A., Emmenegger, C., Kimmich, S., and Hutchins, E. (2012). "Let's look at the cockpit: exploring mobile eye-tracking for observational research on the flight deck," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 107–114.

Wilson, R. A. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences: Cognition*. New York, NY: Cambridge University Press.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 01 March 2023
DOI 10.3389/frai.2022.961595

OPEN ACCESS

EDITED BY
Dietrich Albert,
University of Graz, Austria

REVIEWED BY
Zhao Fan,
Central China Normal University, China
Jianlong Zhou,
University of Technology Sydney, Australia

*CORRESPONDENCE
Philip J. Kellman
✉ kellman@cognet.ucla.edu

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 04 June 2022
ACCEPTED 23 December 2022
PUBLISHED 01 March 2023

CITATION
Baker N, Garrigan P, Phillips A and Kellman PJ (2023) Configural relations in humans and deep convolutional neural networks. *Front. Artif. Intell.* 5:961595. doi: 10.3389/frai.2022.961595

Configural relations in humans and deep convolutional neural networks

Nicholas Baker[1], Patrick Garrigan[2], Austin Phillips[3] and Philip J. Kellman[3]*

[1]Department of Psychology, Loyola University Chicago, Chicago, IL, United States, [2]Department of Psychology, Saint Joseph's University, Philadelphia, PA, United States, [3]UCLA Human Perception Laboratory, Department of Psychology, University of California, Los Angeles, Los Angeles, CA, United States

Deep convolutional neural networks (DCNNs) have attracted considerable interest as useful devices and as possible windows into understanding perception and cognition in biological systems. In earlier work, we showed that DCNNs differ dramatically from human perceivers in that they have no sensitivity to global object shape. Here, we investigated whether those findings are symptomatic of broader limitations of DCNNs regarding the use of relations. We tested learning and generalization of DCNNs (AlexNet and ResNet-50) for several relations involving objects. One involved classifying two shapes in an otherwise empty field as same or different. Another involved enclosure. Every display contained a closed figure among contour noise fragments and one dot; correct responding depended on whether the dot was inside or outside the figure. The third relation we tested involved a classification that depended on which of two polygons had more sides. One polygon always contained a dot, and correct classification of each display depended on whether the polygon with the dot had a greater number of sides. We used DCNNs that had been trained on the ImageNet database, and we used both restricted and unrestricted transfer learning (connection weights at all layers could change with training). For the same-different experiment, there was little restricted transfer learning (82.2%). Generalization tests showed near chance performance for new shapes. Results for enclosure were at chance for restricted transfer learning and somewhat better for unrestricted (74%). Generalization with two new kinds of shapes showed reduced but above-chance performance (≈66%). Follow-up studies indicated that the networks did not access the enclosure relation in their responses. For the relation of more or fewer sides of polygons, DCNNs showed successful learning with polygons having 3–5 sides under unrestricted transfer learning, but showed chance performance in generalization tests with polygons having 6–10 sides. Experiments with human observers showed learning from relatively few examples of all of the relations tested and complete generalization of relational learning to new stimuli. These results using several different relations suggest that DCNNs have crucial limitations that derive from their lack of computations involving abstraction and relational processing of the sort that are fundamental in human perception.

KEYWORDS

perception of relations, deep convolutional neural networks, DCNNs, deep learning, abstract relations, visual relations, shape perception, abstract representation

1. Introduction

The perception of objects, spatial layouts, and events are crucial tasks of intelligent systems, both biological and artificial. For these tasks, information in reflected light affords the richest information. Differences in material substances' absorption and reflection of light carry information about boundaries and shapes of objects and surfaces, as well as their spatial location and relations, textures, and material properties. The concentration of research effort on vision in human and artificial systems is no accident, given the detailed information available in reflected light, its spatial and temporal precision, and its availability at a considerable distance from objects and events themselves.

In human vision, research has identified specialized processes and neural mechanisms that contribute to visual perception and representation of objects, spatial layout, motion, and events. Among these are processes that separate figure from ground and determine border ownership (Rubin, 1915/1958; Koffka, 1935; Driver and Baylis, 1996; Zhou et al., 2000), detect complete objects despite fragmentation due to occlusion or camouflage (Michotte et al., 1964; Kanizsa, 1979; Kellman and Shipley, 1991; Kellman and Fuchser, in press), represent the shapes of contours, objects, and surfaces (Wallach and O'Connell, 1953; Ullman, 1979; Marr, 1982; Biederman, 1987; Lloyd-Jones and Luckhurst, 2002; Pizlo, 2008; Elder and Velisavljević, 2009; Baker and Kellman, 2021), determine the direction of motion (Adelson and Movshon, 1982), and use relational information to perceive events (Michotte, 1954; Johansson, 1978). All of these processes appear to involve computational processes and dedicated neural machinery specialized to extract and represent important structural properties of scenes and events.

A consistent hallmark of these and other aspects of human visual processing is the importance of relations. Relations are crucially involved in visual perception in two related but separable ways. First, capturing important properties of the world involves relational information in the optical input and perceptual mechanisms that can extract it. Relevant relations as stimuli for vision often involve considerable complexity (Johansson, 1978; Gibson, 1979; Ullman, 1979; Marr, 1982; Palmer et al., 2006; Baker and Kellman, 2018). Second, the outputs of perception involve explicit representations of relational properties—relations across space, such as shape or arrangement (Koffka, 1935; Baker and Kellman, 2018), or properties based on patterns across time, such as causality or social intention (Heider and Simmel, 1944; Michotte, 1954; Scholl and Tremoulet, 2000). Evidence indicates the abstract nature of these and other perceptual representations (e.g., Izard et al., 2009; Hummel, 2011; Baker and Kellman, 2018). The representation of relational properties in the output allows perceptual descriptions to subserve a wide variety of tasks and to connect naturally to thought, action, and learning (Gibson, 1969; Garrigan and Kellman, 2008; Klatzky et al., 2008; Kellman and Massey, 2013).

Efforts in artificial vision have sought to develop algorithms for extraction of information that might produce explicit representations of contours, surfaces, spatial layout, objects, and shape (Marr, 1982). For object recognition, these efforts have led to proposals for solving the relevant computational tasks explicitly using information about shape (Bergevin and Levine, 1993; Belongie et al., 2002; Pizlo, 2008; Rezanejad and Siddiqi, 2013), local texture patterns (Lowe, 1999), or surface feature segmentation (Shi and Malik, 2000; Shotton et al., 2009).

Although these efforts have yielded important progress, they have been overshadowed in recent years by results from a wholly different approach: deep convolutional neural networks (DCNNs). DCNN architectures have many applications, but one clear focus, and area of conspicuous success, is in image classification. In DCNNs, object recognition is not based on explicitly encoded contours, surfaces, or shapes of objects present in images (Krizhevsky et al., 2012). Instead, the networks learn to accurately classify many images depicting various object categories from the weighted combination of the responses of many small, local filters, the responses of which are themselves learned.

The successes of deep networks in object recognition have led to research questions flowing in the opposite direction from many earlier efforts. Rather than starting with biological vision phenomena, such as segmentation of figure from ground or completion of partly occluded objects, and attempting to construct computer vision models to perform these tasks, many researchers are currently investigating similarities between deep networks trained for object recognition and the human visual system. Node activity in intermediate layers of deep networks correlates with activity of cell populations in V4 (Pospisil et al., 2018) and some deep networks have been found to be predictive of cell populations in IT (Yamins et al., 2014). Deep networks trained for object recognition also appear to predict human behavior in judging the similarity between objects (Peterson et al., 2016), the memorability of objects (Dubey et al., 2015), and the saliency of regions in an image (Kümmerer et al., 2014).

At the same time, other research has suggested that deep learning approaches have deep limitations. These limitations are being studied in terms of the applicability of deep learning systems as models of biological processing but also regarding their impact in applications to consequential real-world tasks. Ultimately, such inquiries may help to determine both the ways in which the characteristics of deep learning networks are embodied in aspects of biological vision and ways in which deep learning approaches can be enhanced by incorporating specialized adaptations that are evident in biological systems.

In earlier work, we reported that DCNNs that successfully classify objects differ from human perceivers in their access to and use of shape (Baker et al., 2018). Kubilius et al. (2016) had tested shape as a cue for recognition and found that DCNNs can classify silhouettes with about 40% accuracy and showed sensitivity to non-accidental features of objects [e.g., parallel vs. converging edges (Biederman, 1987)]. In our research, we showed that DCNNs showed a clear lack of sensitivity to global shape information. This conclusion rested on multiple, converging tests. When texture and shape conflicted (as in a teapot with golf ball texture), the networks classified based on texture; glass ornaments readily recognizable by humans as animals or objects were poorly classified by DCNNs; DCNNs showed poor performance in classifying silhouettes of animals, and they showed no ability to correctly classify outline shapes (Baker et al., 2018). Examining error patterns led us to suggest a distinction between local contour features and more global shape. DCNNs clearly access the former but seem to have no access to the latter. We tested this hypothesis with silhouettes of objects that DCNNs had correctly classified, altered in two different ways in separate experiments.

In one, we scrambled the spatial relations between object parts to destroy their global shape features while preserving many of the local edge properties present in the original stimulus. In the second, we preserved global shape but altered local edge features by adding serrations to the bounding contours of objects. Although human recognition of part-scrambled objects was highly disrupted, DCNN responses were little affected by scrambling. In contrast, the use of local serrated edges to define overall shape had little effect on human classifications but completely disrupted the network's classification of objects (Baker et al., 2018).

Subsequent work provided further evidence that DCNNs have little or no sensitivity to global shape. Baker et al. (2020b) found that networks they trained to discriminate squares and circles would consistently classify as circles squares whose edges were comprised of concatenations of curved elements. Similarly, circular patterns made from concatenations of small corner elements were classified as squares. These results were relatively consistent across a variety of DCNNs (AlexNet, VGG-19, and ResNet-50), and for both restricted and unrestricted transfer learning (Baker et al., 2020b).

These and other results pose clear contrasts with research on human visual perception, in which shape is the primary determinant of object recognition (Biederman and Ju, 1988; Lloyd-Jones and Luckhurst, 2002; Elder and Velisavljević, 2009). Shape is represented even when it must be abstracted from disconnected stimulus elements (Baker and Kellman, 2018). In fact, the specific, directly accessible local features from which shape is extracted are often not encoded in any durable representation (Baker and Kellman, 2018) and may in many cases be represented as statistical summaries rather than precise records of features in particular positions (Baker and Kellman, in press).

1.1. Motivation of the present research

It might be natural to interpret the limitations of DCNNs with regard to global shape as deriving from the absence in these networks of specialized shape extraction and representational processes that have evolved and proven useful in human vision. Although we believe aspects of that point of view are likely correct, we wondered whether the limitations in capturing shape relations in DCNNs might be indicative of a more general limitation regarding relations.

A basic reason for supposing that DCNNs might have a general limitation with regard to relations involves the convolution operation at the heart of much of DCNN processing. Convolution applied to an image input is inherently a local process and a literal process. The output of a convolution operator at the location of its center is the weighted sum of image values of intensity in a neighborhood of locations around the center. At later layers, convolution may be applied to the values obtained by a prior convolution operation or some kind of pooling operation, such as max pooling, which reduces the size of the array by assigning to larger neighborhoods the maximum value of operator outputs in that region. There is little doubt that these operations have high utility and flexibility. The convolutional kernels that develop through learning can assume a vast variety of forms. Likewise, one or more fully connected layers in a DCNN can allow the development, through changes of weights in training, of sensitivity to a wide variety of relations between even spatially separated locations. DCNNs can theoretically capture an enormous number of potential relations in images, many of which would defy easy verbal description by humans and would never be designed in a priori attempts to capture important properties.

Yet not all relations are created equal. There may still be important limitations regarding most DCNNs and relations. In particular, relations that require explicit representation or abstraction may be problematic. This idea would fit with previously discovered limitations regarding shape. As emphasized in classic work by Gestalt psychologists (e.g., Koffka, 1935), shape is an abstract relational notion. A square may be made of small green dots in particular locations, but neither relations defined over green dots nor specific locations are intrinsic to the idea of squareness. Any tokens will do to define the spatial positions of parts of a square, and the particular spatial positions do not matter. In the end, being a square is neither local in requiring elements to occur in a particular place nor literal in requiring green dots or any other specific kind of local stimulus properties. What is crucial to squareness is the spatial relations of the elements, not a concatenation of the pixel values of the elements themselves. Research on human shape perception provides evidence for the primacy of abstract, symbolic representations (Baker et al., 2020a). With their roots in convolution operations, DCNNs excel in leveraging relations of a concrete sort, involving specific local features and color values, but they may lack mechanisms to extract spatial relations, abstracting over the concrete properties of elements (Greff et al., 2020); learning of this sort may require dedicated computational machinery that separates the representation of relations and their arguments (Hummel, 2011).

Some recent work has tested the capabilities of DCNNs to learn visual relations, with particular consideration of their capacities to solve same-different problems. Findings from these investigations indicate that basic DCNNs, as well as some older well-established DCNN architectures (e.g., AlexNet, VGG, LeNet, and GoogLeNet) struggle with same-different tasks, while some newer networks (e.g., ResNets and DenseNets) perform better (Stabinger et al., 2016; Kim et al., 2018; Messina et al., 2021). However, subsequent work by Puebla and Bowers (2021) found that ResNet-50, a 50-layer, enhanced version of earlier ResNets, failed to generalize same-different relations when test images were dissimilar from training images at the pixel level. So far, there is no compelling evidence that deep networks learn relations such that they can apply them to new displays.

In the present work, we aimed to test a variety of relations in visual displays that human perceivers would notice and learn with little effort from a small number of examples, and generalize accurately to new examples. We attempted to replicate and further explore the same-different relation in DCNNs and test two new relations to look at overall characteristics of DCNNs and relational generalization, while using human performance as a comparison.

1.2. Plan of the experiments

In Experiment 1, we investigated the learning and generalization of same-different relations in pairs of displayed objects. In Experiment 2, we investigated the relationship of enclosure; each display had a dot that fell either inside or outside of the only closed figure in the display. In Experiment 3, we tested a relationship between color and an object property. Both deep networks and humans were trained and tested in a two-alternative categorization task with displays having two polygons. Whether the display fell

into one category or the other depended on whether the polygon with a red dot inside it had a greater or fewer number of sides than the other polygon. For each relation, we trained DCNNs using restricted and unrestricted transfer learning in separate studies. After the completion of training, we tested for generalization to members of the training set withheld during training. We then tested for generalization with new displays that differed in some object characteristics but embodied the same relation that had been the focus of training. In parallel, we also carried out studies with human observers to assess whether the relation in question could be quickly discovered and used for classification and generalization.

2. Learning same-different relations

2.1. Experiment 1a: Same-different training

We first tested DCNNs' ability to learn same-different classifications. In this task, we placed two novel, closed contours in a single image and tasked the network with learning to produce a "Same" response when the shapes of both contours were the same as each other, and a "Different" response otherwise. The same-different task would be learnable if DCNNs can obtain a feature description of two objects individually within an image and then make a classification decision based on the relation between these two feature descriptions. This differs from standard classification tasks in which the feature descriptions themselves, not the relations between feature descriptions, are pertinent to the network's classification decision.

2.1.1. Method

2.1.1.1. Network

All tests were conducted on AlexNet (Krizhevsky et al., 2012) and ResNet-50 (He et al., 2016), pre-trained on ImageNet (Deng et al., 2009). AlexNet is a high-performing DCNN with relatively few convolutional layers, while ResNet-50 is a much deeper network that represents the current state-of-the-art in feedforward DCNNs.

2.1.1.2. Training data

In each of the experiments presented in this paper, artificial images were generated so that categorization by a DCNN required sensitivity to the relationship being tested. Artificial images, rather than digital images of natural scenes, were used for two reasons. First, it would be difficult to find sufficient number and variety of natural images, and second, it would be difficult or impossible to assess whether classification was based on the relationship of interest, or some other correlated, non-relational cue.

We generated 20 novel shapes by moving 10 control points toward or away from the center of a circle, then fitting cubic splines between these control points (see Baker and Kellman, 2018). Training data consisted of images in which one of the 20 shapes appeared twice in the image ("Same" trials) and in which two of the 20 shapes appeared in the image, once each ("Different" trials). In order to prevent overfitting, we placed both shapes in random positions within the image frame with constraints so that the two contours did not overlap and did not touch the image boundary. Each shape was randomly assigned one of 10 sizes, which varied between 20% and 30% of the length of the image frame along the shape's longest dimension. In total, we created 10,000 "Same" and 10,000 "Different" training images. Figure 1 shows some sample "Same" and "Different" images used in training.

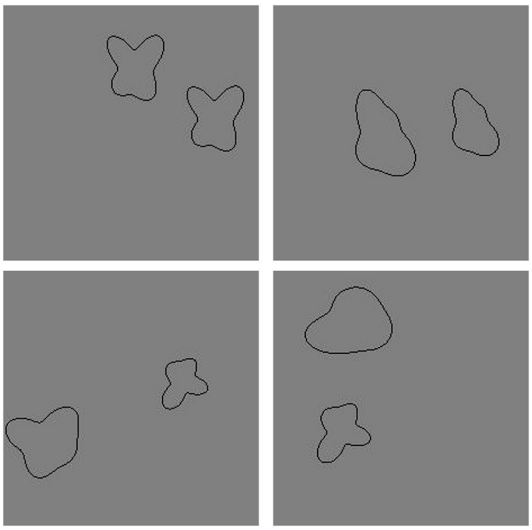

FIGURE 1
Sample images used during training in Experiment 1a. **(Top)** Two "Same" images. **(Bottom)** Two "Different" images.

2.1.1.3. Training

In order to assess whether DCNNs could learn the same-different relation, we used two different types of transfer learning on an ImageNet-trained AlexNet architecture. In one, we froze all connection weights between convolutional layers in AlexNet, allowing only the last set of connection weights between the penultimate layer and the classification layer to update. We call this restricted transfer learning. Restricted transfer learning tests whether a sensitivity is already latently present from ImageNet training, because the output or decision layer of a network is necessarily based on some weighted combination of the activation of the 4,096 nodes in the penultimate layer. If coding sufficient to detect the presence of two objects of the same shape in a display had evolved in prior training of a DCNN to classify objects, then restricted transfer learning might learn to perform accurately this two-choice discrimination by discovering appropriate combinations of node activations in the penultimate layer.

The second form of transfer learning, unrestricted transfer learning, also begins with a pre-trained network, but allows connection weights at all layers to update during the learning of the new classification task. Unrestricted transfer learning assesses DCNNs' more general capability of obtaining a particular sensitivity, regardless of whether that sensitivity was previously present or not.

We trained with a minibatch size of 32 and an initial learning rate of 1×10^{-5}. We used 80% of our training data for training and withheld 20% as a validation set. We trained for up to 10 epochs or until error rates on the validation set increased six consecutive times.

For ResNet-50, based on our findings with AlexNet, we used only unrestricted transfer learning. The training data were identical to the data used to train AlexNet. We used a batch size of 50 and an initial learning rate of 1×10^{-3}. We began by training ResNet-50 for 10 epochs and then did a second training experiment with 70 epochs.

2.1.2. Results

With restricted transfer learning, AlexNet reached criterion after three epochs. Although error rates had increased six consecutive times on the validation set, the network's final classification accuracy showed no evidence of sensitivity to the same-different relation. Performance on the validation set was 54.4%, close to chance performance for the binary classification task, and similar to accuracy levels shown at the end of training. These results suggest that the same-different relation is not something acquired or naturally encoded during training on the ImageNet dataset.

With unrestricted transfer learning, AlexNet reached criterion after 10 epochs. Compared to other transfer learning tasks that do not require a relational comparison (Baker et al., 2020b), learning for the same-different task was both slower and weaker, but the network did eventually improve to 82.2% performance on the validation set, well above chance responding.

After 10 epochs, ResNet-50 did not achieve above-chance classification on the validation set (mean accuracy = 49.7% on the validation set). To assess whether the network simply needed more training iterations to achieve accurate classification, we repeated training with 70 epochs. More extended training produced only a modest improvement in classification accuracy, from 49.7 to 56.0%.

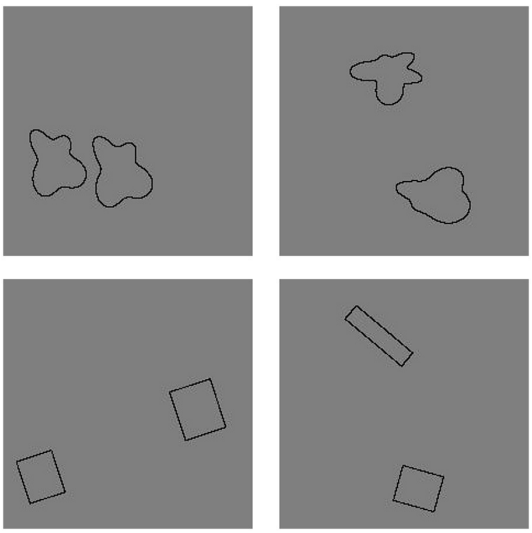

FIGURE 2
Example generalization test images in Experiment 1b. **(Top)** A "Same" and a "Different" image for the first generalization test. **(Bottom)** A "Same" and a "Different" image from the second generalization test.

2.2. Experiment 1b: Generalization following unrestricted transfer learning

When all connection weights were allowed to update, AlexNet achieved well above chance performance on the same-different task. Our key question here, however, involved what was learned? Did the network learn to attach certain responses to certain images, allowing it to achieve above-chance performance? Or did it come to classify based on detecting sameness or difference between two objects in each display? To test whether the network had learned the abstract "Same" relationship or whether its accurate responses were specific to the shapes we used during training, we generated new images with pairs of shapes that included new shapes qualitatively similar to the shapes used in training, and shapes qualitatively different from those used in training. If the network had come to use the abstract relation, its performance should generalize to new shape pairs.

2.2.1. Method

We used two generalization tests to assess the networks' generalization of the same-different rule. First, we generated 30 new "Same" and 30 new "Different" shapes using the same algorithm previously used to generate the shapes used in training. As in training, the pairs of shapes were given a random size and position in the image frame with constraints to prevent them from overlapping and extending out of the frame.

We also wanted to test the networks' generalization to the same-different relation using dissimilar shapes. For this test, we used pairs of rectangles. We generated images with two rectangles. The ratio of the minor to principal axis of the rectangles was randomized and varied from 0.08:1 to 1:1. Both rectangles were placed in the image with random size and position. In the "Same" trials, both rectangles in the image had the same aspect ratio and differed only by size and position. In the "Different" trials, the two rectangles differed in aspect ratio as well as by rigid 2D transformations. We generated 30 "Same" and 30 "Different" rectangle pair stimuli. Examples of images from both generalization tests are shown in Figure 2.

We tested both AlexNet and ResNet-50 trained with unrestricted transfer learning on both new sets of stimuli. Because the networks trained with restricted transfer learning never achieved above-chance performance on the validation set, there was no reason to apply the generalization tests to it.

2.2.2. Results

AlexNet's performance was poor in both generalization tests. For the test in which new shapes were generated from the same method as in training, network performance fell from 82% to 58%. For the test with rectangles, performance fell to 50%, with the network classifying all pairs of rectangles as "Same."

For ResNet, performance was already poor but fell fully to chance on the generalization tests. The network trained with unrestricted transfer learning classified 45% of the new shape stimuli correctly and 50% of the rectangle stimuli correctly.

2.3. Experiment 1c: Comparison with humans

The results of our transfer learning experiment on DCNNs suggests they have little ability to use the abstract same-different relation in order to classify images. Humans' registration of same-different relations in perceptual arrays is rapid and automatic (Donderi and Zelnicker, 1969). However, it is possible that our specific paradigm does not elicit perception of sameness/difference in humans. If this were true, then the lack of generalization we saw in DCNNs might not point to a difference in perceptual processing between networks and humans. We tested this by conducting the same experiment we used on DCNNs on human participants.

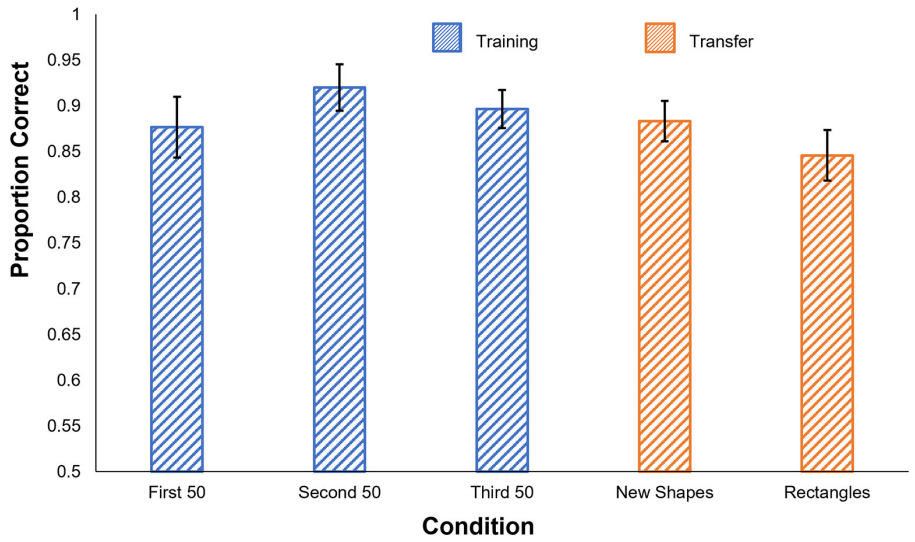

FIGURE 3
Human results in Experiment 1c. Proportion correct is shown by condition. Blue: performance in the training phase, separated into 50-trial blocks. Orange: performance on the generalization tests. Error bars show ± one standard error of the mean.

2.3.1. Method

2.3.1.1. Participants

Six undergraduates (two female, four male, $M_{age} = 21.0$) from Loyola University participated in this experiment as lab researchers. All participants were naive to the purpose of the experiment before completing it.

2.3.1.2. Design

The experiment consisted of a learning phase (150 trials) and two generalization phases (40 trials each). The first generalization phase tested whether classification based on sameness/difference would generalize after learning to new shapes generated in the same way as shapes in the learning phase. The second generalization phase tested pairs of rectangles having the same or different aspect ratios.

2.3.1.3. Stimuli

All stimuli used in the human experiment were taken directly from images used to train or test AlexNet in our DCNN experiment. For the learning phase, we randomly selected 150 (75 same, 75 different) images used during transfer learning. For the generalization tests, we randomly selected 20 same and 20 different images from the same tests used on DCNNs.

2.3.1.4. Procedure

At the beginning of the experiment, participants were told that they would be classifying images into two categories but that they would not be told what defined the two categories. Their task was to use accuracy feedback to discover how to classify images.

During the training phase, participants were shown an image on the screen for 500 ms, after which they were asked whether the previous image belonged to Category 1 or Category 2. After responding, participants were told whether they were correct or incorrect and given the correct classification for the previous image. The image was not shown again during feedback.

Following the training phase, participants completed two generalization tests. They received no feedback during the generalization phases but were told to continue using the same criteria they had adopted during the training phase. In the first generalization test, participants were shown images with the same types of shapes they saw during training, but the actual shapes were different. In the second generalization test, participants were shown images of rectangles with the same or different aspect ratios.

2.3.1.5. Dependent measures and analysis

To assess learning in the learning phase, we separated trials into three 50-trial blocks corresponding to the first, middle, and last third of trials. Because we hypothesized that humans would readily perceive abstract relations such as same vs. different, we predicted that by the second 50-trial block, participants would have learned the rule for categorizing images and should respond correctly for nearly every image.

To assess learning in the testing phases, we simply measured participants' proportion correct and compared their performance on the generalization tests with chance performance and with performance on the final block of the learning phase.

2.3.2. Results

The results of the human experiment are shown in Figure 3. Participants performed very well even in the first 50-trial training block and reached ~90% in each of the last two blocks. *t*-tests confirmed that participants performed significantly better than chance in all three training blocks [1st block: $t_{(5)} = 11.33$, $p < 0.001$; 2nd block: $t_{(5)} = 16.60$, $p < 0.001$; 3rd block: $t_{(5)} = 18.96$, $p < 0.001$].

2.3.2.1. Generalization

Participants' accuracy remained high in both generalization tests, significantly exceeding chance levels [New Shapes: $t_{(5)} = 17.39$, $p < 0.001$; Rectangles: $t_{(5)} = 12.48$, $p < 0.001$]. Performance levels also did not significantly differ between the last 50 trials of the training phase and either of the generalization tests [New Shapes: $t_{(5)} = 1.10$, $p = 0.32$; Rectangles: $t_{(5)} = 1.65$, $p = 0.16$].

2.4. Discussion, Experiments 1a–c

Research has shown that DCNNs' recognition of objects is primarily driven by texture information, rather than the shape information preferentially used by humans (Baker et al., 2018; Geirhos et al., 2018). Whereas textures and local shape features are composed of locally defined elements, global shape involves relationships among spatially separated parts of object boundaries. Considerable evidence indicates that this more global notion of shape, as opposed to local shape features, is not accessible to DCNNs, even when texture is made non-informative for classification (Baker et al., 2018, 2020b). When texture information is unavailable to DCNNs, they may still achieve above-chance classification accuracy using local contour cues, but not more global features of shape (Baker et al., 2018, 2020b).

We hypothesized that DCNNs' insensitivity to shape may be caused by a more general insensitivity to relational information. To test this idea, we presented the network with a classification task with class type defined by the relation "Same-Different." With restricted transfer learning, there was no indication that the network could learn this classification. This result is perhaps not surprising, since we did not expect that a DCNN trained for image classification would have sensitivity to global shape. Interestingly, however, with unrestricted transfer, AlexNet did learn to classify the trained shape pairs as same or different (independent of their sizes and positions), but the learning was specific to the trained shapes. Performance was near chance for novel shapes, created through the same generative procedure, and for rectangles. Humans trained with the same shapes showed robust generalization in both cases.

The human visual system is highly flexible, able to represent visual information differently depending on task and stimulus constraints. In numerical cognition research, humans can flexibly switch between perceiving individual objects (Piazza et al., 2011; Cheng et al., 2021), ratios between object groups (He et al., 2009), and objects as a texture field (Burr et al., 2017), depending on stimulus constraints. Similarly, in shape perception, humans can flexibly switch between more local and more global features of a shape (Navon, 1977; Kimchi, 1998; Bell et al., 2007), although the global percept is stronger in many cases. In contrast, DCNNs appear to be much less flexible, making their classifications based only on a small subset of the visual information considered by humans.

The inability of DCNNs to acquire and generalize the same-different relation here is not a finding that arises predictably from prior evidence of the lack of global shape encoding in DCNNs. As mentioned, using unrestricted transfer learning, we did see evidence of acquisition of above-chance performance with the training set. More conceptually, the initial same-different learning task and the first generalization task we posed to the networks could have been accomplished to a high degree of accuracy by use of local shape features without global shape encoding. The notion of same-different can just as well apply to unstructured collections of local features as to global shape. To give one example, in the amoeboid figures, similarities in signs of local curvatures could be informative in determining sameness (in contrast, the rectangles used in the second generalization test may have fewer distinguishing local features; hence all pairs were classified as "Same"). Where available, as in the amoeboid figures, local shape information could have supported the above chance performance on the training set in unrestricted transfer learning. The crucial result regarding relations, however, is that whatever was used to produce correct "Same" and "Different" responses in training showed little or no generalization to new shapes, indicating that whatever was learned, it was not the abstract relation of sameness.

The idea that (somewhat) successful same-different classification observed in training (but not in generalization) was based, not on the relationship same-different, but on the development of sensitivity to the co-occurrence of local features across specific shape pairs aligns with recent work by Puebla and Bowers (2021), who found that DCNNs could only generalize the same-different relation to stimuli that matched training data at a pixel level. The result is impressive, given that the positions and sizes of the shapes in each pair were varied independently, and it underscores the massive capacity for DCNNs to map many different feature combinations onto discrete categories.

The fact that learning did not generalize beyond the trained set, though, as evidenced by the lack of generalization to novel shapes, similarly underscores a key limitation of the operation of these DCNNs. One would expect that, following training, humans could perform this classification on a limitless number of novel shape pairs, provided the shapes themselves were not too complicated or the differences between members of the pairs too subtle. With increased complexity and sufficient training data, a network with this type of architecture would likely be able to learn to successfully classify a larger variety of shape pairs (up to limitations imposed by the vanishing gradient problem), but it would still only be able to classify novel shape pairs to the extent that they resembled pairs in the training data.

In contrast, ResNet-50 never achieved better than near-chance accuracy on the same-different task, even with unrestricted transfer learning and many training epochs. It is puzzling that the deeper network performed worse than AlexNet. Based on AlexNet's poor performance on the generalization tests, it seems likely that whatever rule it was using to perform above chance in training was highly stimulus-specific, not an abstract visual relation. One difference between AlexNet and ResNet is that AlexNet has two fully connected layers between the convolutional layers and the decision layer whereas ResNet has only convolutional layers. These fully connected layers might be important for relating widely spaced features in an image, a process that may be important for the non-abstract comparison furnishing above-chance performance in the training data for AlexNet.

Issues relating to limitations of connectionist networks in capturing or representing abstract relations have been recognized for some time (e.g., Hummel, 2011). The architecture of DCNNs, although more powerful than earlier connectionist approaches, due to both hardware advances (e.g., leveraging GPUs for greater processing power, more memory) as well as algorithmic changes (convolutional layers, skip connections, pooling, etc.), share this same limitation with their ancestors. That said, a more sophisticated network might be able to exhibit some processing of relations, despite these limitations, within a restricted domain. In fact, recent evidence shows that activity in intermediate layers consistent with Weber's Law and sensitivity to the relative sizes of objects, properties that appear to involve simple spatial relations, emerges spontaneously in DCNNs trained for object recognition (Jacob et al., 2021). Our results show, however, that even in this one restricted domain (same-different

FIGURE 4
Example training images for Experiment 2a. **(Top)** Two "Inside" images. **(Bottom)** Two "Outside" images. Category membership was determined by the position of the small red square, either inside or outside of a closed contour.

shape judgments on closed, 2-D contour stimuli), there was little evidence the network could learn to classify based on relational processing outside of the trained set.

It is possible that DCNNs could perform better for other sorts of relational tasks. In Experiments 1a–c, we tested "Same-Different" shape classification performance while allowing for changes in the sizes and positions of the shapes in each comparison pair. Same-different shape classification, while a very intuitive task for people, might be a particularly challenging case for DCNNs. While the task was made easier by not including rotations between the members of a "Same" pair, the network still needed to handle considerable variability both in the shapes themselves and their presentation (i.e., position and size), and to learn to distinguish the features and their relations within a single shape from those between shapes. In Experiments 2 and 3, we consider other relational properties.

3. Learning an enclosure relation

In Experiments 2a–b, we investigated a relational property that is perhaps a bit more constrained than abstracting sameness or difference and applying those to novel shapes. We tested the relation of enclosure, specifically, whether a small, locally-identifiable object (a red dot) was inside or outside of a closed contour.

3.1. Experiment 2a: Enclosure training

A contour is closed if it has no gaps and its curvature integrates to 360°. In humans, contour closure is a salient cue; it confers perceptual advantages in detection (Kovacs and Julesz, 1993), search (Elder and Zucker, 1993), and recognition tasks (Garrigan, 2012). Experiment 2 specifically aimed to test whether humans and DCNNs can learn to classify images based on an abstract relation between a dot and a closed contour. In one category of images ("Inside"), the dot is within a region is surrounded by a closed contour while in the other category ("Outside") the dot is outside the region surrounded by the closed contour. Each display had only one closed contour present, along with open contours as noise fragments to eliminate certain possible correlates of enclosure that might otherwise allow DCNNs to perform successfully without detecting the enclosure relation.

3.1.1. Method

3.1.1.1. Network

As in Experiment 1, all tests were conducted on AlexNet and ResNet-50 pre-trained on ImageNet.

3.1.1.2. Training data

For both image categories, we generated a closed contour by moving 10 control points toward or away from the center of a circle and fitting cubic splines between the control points. The shapes were sized so that the greatest distance between two vertical or two horizontal points was between 16.7% and 33.3% of the length of the image frame. The contour was randomly positioned in a 227 × 227 pixel image with the constraint that the whole contour must be within the image frame.

In addition to the closed contour, we added 22 unclosed contour fragments to the image in random positions. The unclosed contour fragments were generated by forming contours in exactly the same way as the closed contour, but selecting only 25–50% of the full contour.

For "Inside" images, we placed a red probe dot in a random position within the closed contour with the constraint that it could not touch the closed contour's border. For "Outside" images, a red probe dot was placed somewhere in the image outside of the region enclosed by the closed contour's border. We constrained the positions of the probe dots in the "Outside" images to be at least 23 pixels away from edges of the full display because these probe positions were unlikely for "Inside" images. We generated 1,000 "Inside" and 1,000 "Outside" images to use as training data for the DCNN. Sample images are shown in Figure 4.

3.1.1.3. Training

As in Experiment 1, we trained AlexNet using both restricted and unrestricted transfer learning. We trained with 90% of our training data, withholding 10% as a validation set. All other training parameters were the same as in Experiment 1. Training concluded after 10 epochs or after the error rate on the validation set increased in six consecutive trials.

Training of ResNet-50 also followed Experiment 1. We trained for 10 epochs using unrestricted transfer learning.

3.1.2. Results

Training with restricted transfer learning ended after eight epochs. The network's accuracy on the validation set was 51.0% after training, around chance levels for a binary classification task. As in Experiment 1, the features learned through ImageNet training do not appear to be usable for the inside/outside task.

Unrestricted transfer learning ended after 10 epochs, with an accuracy of 74.0% on the validation set. These results align with the findings of Experiment 1 and transfer learning in other tasks (Baker et al., 2020b) in that performance was better with unrestricted transfer learning.

Unlike in Experiment 1 where ResNet-50 performed much worse than AlexNet in training, the deeper network performed significantly

better in the inside/outside task. Performance reached 99.8% on the validation set after 10 training epochs.

3.2. Experiment 2b: Generalization to other enclosure tasks

Had the network learned the abstract enclosure relation? In order to test this, we generated new stimuli in which the inside/outside relation was unchanged, but certain irrelevant image properties differed from the network's training data. The first two generalization tests we conducted tested whether changing contour properties of the closed shape and the open contour fragments would affect the network's classification performance. First, we adjusted a parameter in our generative method for producing shapes to see whether the network generalized. Next, we changed the contours from amoeboids to squares and parts of squares. Our final generalization test evaluated a specific hypothesis that the network's above-chance responding was based on probe dot's proximity to the closed contour boundary, not enclosure of the probe dot. We hypothesized that if this were true, then by making the contour bigger, network performance would fall.

3.2.1. Method

In our first generalization test, we generated shape contours by fitting cubic splines through 16 control points moved away from a circle's boundary rather than the 10 control points used in our training data. Both the closed contour and the contour fragments were generated with 16 control points instead of 10. All other parameters were the same as in the training data. We generated 30 "Inside" and 30 "Outside" images with the new parameter in our generative method.

In our second generalization test, we generated shape contours with squares instead of amoeboid shapes produced by fitting cubic splines through control points. The squares were constrained to be of approximately the same size as the shapes generated in training. As in the training stimuli, open contour fragments were added by randomly selecting 25–50% of square contours that were otherwise matched with the closed contour. We generated 30 "Inside" and 30 "Outside" images with square contours.

In our final generalization test, we kept all parameters the same as in training except that we made the closed shape contour significantly larger to increase the distance between the probe dot and the boundary in "Inside" stimuli. We changed the closed shape's size so that the longest horizontal or vertical distance between any two points on the shape's contour was 80% of the length of one side of the image frame rather than 16.67–33.33% as was used in the training data. Sample images for all three generalization tests are shown in Figure 5.

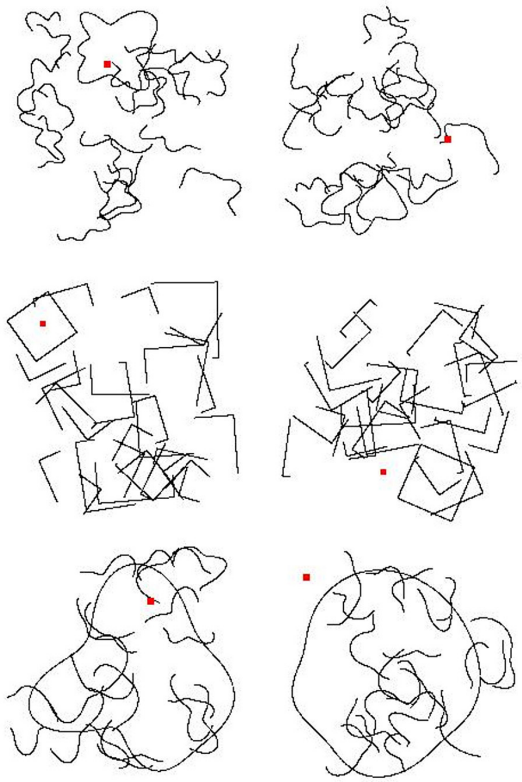

FIGURE 5
Example generalization test images for Experiment 2b. **(Top)** An "Inside" and an "Outside" image for the first generalization test. **(Middle)** An "Inside" and an "Outside" image from the second generalization test. **(Bottom)** An "Inside" and an "Outside" image for the third generalization test. Category membership was determined by the position of the small, red square, either inside or outside of a closed contour.

3.2.2. Results

In all three generalization tests, network performance fell considerably. For the generalization test with 16 control point amoeboids, network performance fell from 74% to 63% for AlexNet and from 99.8% to 76.7% for ResNet-50. For the generalization test with square contours, network performance fell from 74% to 65% for AlexNet and from 99.8% to 59.7% for ResNet-50. For the generalization test with larger contours, network performance fell from 74% to 57% for AlexNet and from 99.8% to 60.0% for ResNet-50.

3.3. Experiment 2c: Comparison with humans

Once again, we found little evidence that the DCNN's above-chance performance in the enclosure task was due to apprehension of the abstract inside/outside relation. Instead, DCNNs appear to be using some kind of combination of cues about where in the image the probe dot is positioned (independent of the location of the closed contour) and the probe dot's distance from contours. In Experiment 2c, we tested whether humans, when exposed to the same training displays as networks, learned to use the abstract inside/outside relation and if the use of this relation produced accurate responding on generalization tests.

3.3.1. Method

3.3.1.1. Participants

Six undergraduate (three female, three male, $M_{age} = 21.0$) from Loyola University participated in this experiment as lab researchers. Five of the six participants were the same as in Experiment 1c. All

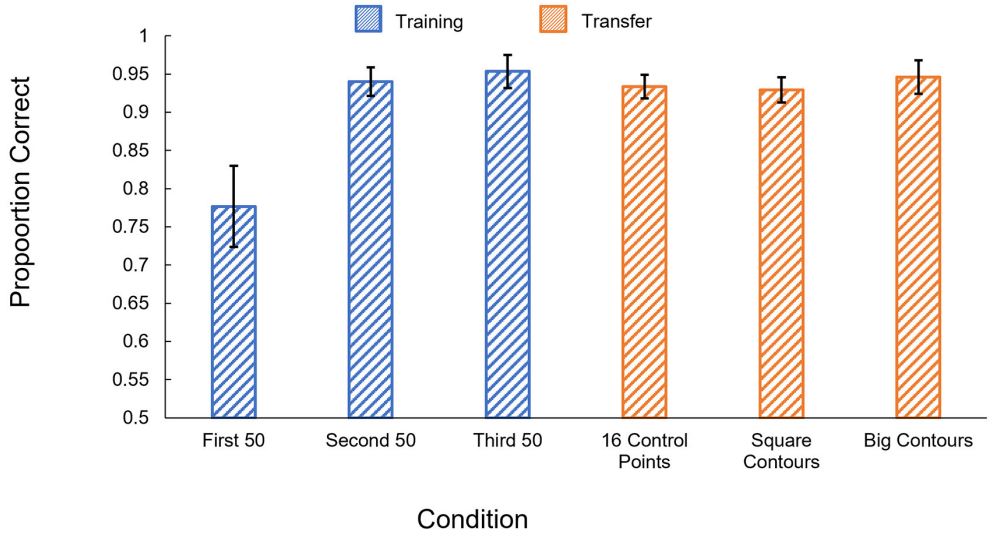

FIGURE 6
Human results in Experiment 2c. Proportion correct is shown by condition. Blue: performance in the training phase, separated into 50-trial blocks. Orange: performance on the generalization tests. Error bars show ± one standard error of the mean.

participants were naive to the purpose of the experiment before completing it.

3.3.1.2. Design

Experiment 2c consisted of a learning phase with 150 trials and three generalization phases with 40 trials each. The three generalization phases were the same as those upon which the DCNNs were tested after transfer learning.

3.3.1.3. Stimuli

All stimuli used in the human experiment were taken directly from images used to train or test the DCNNs in Experiment 2a and 2b. We once again selected 150 (75 same and 75 different) images used during the learning phase and 20 same and 20 different images from the generalization tests used on DCNNs.

3.3.1.4. Procedure

The procedure was the same as Experiment 1c. The only thing that differed was the images used during the learning and generalization phases.

3.3.2. Results

The results of Experiment 2c are shown in Figure 6. Participants performed significantly better in the second block of the learning phase trials than the first [$t_{(5)} = 3.04$, $p = 0.03$], but appear to have reached ceiling by the second block and show little improvement from the second block to the third [$t_{(5)} = 0.54$, $p = 0.61$]. Participants performed significantly better than chance in all three training blocks [1st block: $t_{(5)} = 5.21$, $p = 0.003$; 2nd block: $t_{(5)} = 23.63$, $p < 0.001$; 3rd block: $t_{(5)} = 20.89$, $p < 0.001$].

Participants showed robust generalization in all three of our tests, performing significantly better than chance [16 control points: $t_{(5)} = 28.2$, $p < 0.001$; Square contours: $t_{(5)} = 26.25$, $p < 0.001$; Big contours: $t_{(5)} = 20.44$, $p < 0.001$]. Performance also did not significantly differ from performance in the last block of the learning phase for any of the three generalization tests [16 control points: $t_{(5)} = 0.94$, $p = 0.39$; Square contours: $t_{(5)} = 0.78$, $p = 0.47$; Big contours: $t_{(5)} = 0.27$, $p = 0.80$].

3.4. Experiment 2a-c discussion

As in Experiment 1, the network was able to perform the classification following unrestricted, but not restricted, transfer learning. Unlike Experiment 1, however, the learning did show some generalization to new conditions, including irregular closed contours generated with a modified procedure (63% and 76.7% for AlexNet and ResNet-50, respectively), and closed rectangles (65% and 59.7%, respectively). We suspected, however, that the network was classifying based on a simpler, more local, relationship—the proximity of the probe dot to a part of any contour in the display. This strategy would naturally account for classification performance reliably above chance, but far from perfect.

To test this idea, we had the model perform the inside/outside classification with larger closed contour shapes, creating displays with more locations "Inside" the closed contour that were also distant from the contour itself. Consistent with our hypothesis, network training generalized the least in this condition (57 and 60%, for AlexNet and ResNet-50, respectively). We investigated this idea more directly by examining the pattern of correct and incorrect classifications for a specific image. In Figure 7, for two stimuli (one isolated closed contour and the same closed contour presented among open contour fragments), classification performance is analyzed for all possible probe positions. In both cases, for virtually all probe positions inside the closed contour, AlexNet classified that position as "Inside." The model's behavior for probe positions outside the closed contour, however, provides more insight.

For the isolated contour, most probe positions outside the closed contour were classified as "Inside," and the errors make little sense

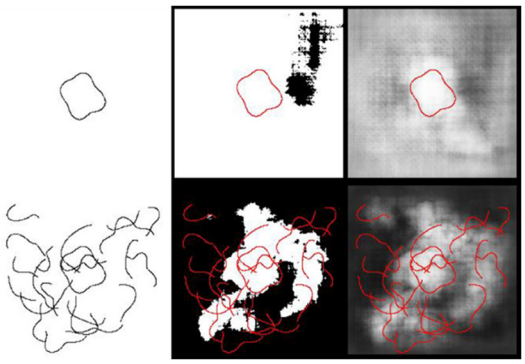

FIGURE 7
"Inside/Outside" classification of sample images following unrestricted transfer learning in Experiment 2. **(Left column)** Two sample input stimuli are shown, including an isolated closed contour **(left, top)** and the same contour presented among contour fragments, replicating training conditions **(left, bottom)**. In both cases, correct classification of the image was "Inside" only if the probe dot (not shown) was presented within the closed contour. **(Middle column)** Model classification results are shown as binary images indicating the model's classification for various positions of the probe, with the input stimulus superimposed (in red). White indicates probe positions classified as "Inside;" black indicates probe positions classified as "Outside." **(Right column)** The same results are shown in the right column, but with *p* (inside) indicated by grayscale values (black = 0.0, white = 1.0).

for a network sensitive to the actual spatial relationship "Inside." For example, it is hard to explain why a network that had learned to encode this relationship would correctly classify a probe in the far upper-right as "Outside," but incorrectly classify probe positions in the three other corners, despite being approximately the same distance from (and not close to) the closed contour. A display with a single, isolated, closed contour, while a useful exploratory tool, is, however, very different from the actual displays used in the training set.

For the closed contour presented among open contour fragments, there was little evidence that proximity of the probe to any contour in the display was driving "Inside" classifications. One might expect errors at probe locations where the contour fragments "almost close," or where the image is particularly cluttered. However, there is little to suggest this is the case. In Figure 7, middle panel in the bottom row, consider the white region in the central, upper region. Correct classifications of "Inside" are represented by the white region approximately centered in the image, bounded by the red contour. The other white regions represent areas misclassified as "Inside." The errors observed in these regions cannot be straightforwardly explained by features of the contour fragments nearby them. In fact, other parts of the image appear, by inspection, to have contour fragments that more closely approximate a closed contour (e.g., on the left side, middle).

While it is unclear what strategy the network uses for achieving above chance classifications in the generalization conditions, comparison with human performance strongly indicates that any relational processing by the network is very different from the strategy employed by humans. Humans learned quickly, achieving near ceiling performance by trials 50–100, suggesting that the inside/outside relationship was salient. Further, complete generalization of learning was observed in all cases.

4. Learning higher-order relations

4.1. Experiment 3a: Network training for higher order relations

In Experiments 1 and 2, we found that humans learn to use perceived abstract relations to categorize images while networks do not. The use of these relations allows human performance to generalize to new stimuli. Networks, although they can learn to classify training stimuli and validation displays similar to the training stimuli, do not extract perceptual relations that allow for generalization of a relation to other kinds of images. Both of the previous experiments tested a simple relation between two image features. For example, in Experiment 1, if the two shapes in the image were the same, the image belonged to the "Same" category. In Experiment 2, if the red dot was within the closed contour, the image belonged to the "Inside" category. These could be called first-order relations because they deal directly with the relation between two properties of an image. A higher order relation would consider a relation between two relations. In Experiment 3, we tested human and DCNNs' ability to classify based on one such higher-order relation.

The images we used in Experiment 3 were displays containing two white polygons on a black background. One of the polygons had a red dot in its center. If the polygon with a red dot had more sides than the polygon without the dot, the image belonged to the "More" category. If the polygon with a red dot had fewer sides than the other, the image belonged to the "Fewer" category. This classification requires the use of a second-order relation because correct responding requires seeing which polygon has more sides, as well as whether that polygon contains the dot.

4.1.1. Method

4.1.1.1. Network

As in Experiments 1 and 2, we trained and tested AlexNet and ResNet-50, pre-trained on the ImageNet database.

4.1.1.2. Training data

Each image in our training data consisted of two polygons with three to five sides. Images were constrained to always include two polygons with a different number of sides. The size of the image was 227 × 227 pixels. Polygons ranged in length from 22 to 42 pixels and in orientation from 0 to 360°. In each image, we placed a red dot at the center of one of the two polygons. We created 10,000 images in which the red dot was at the center of the polygon with more sides ("More" trials) and 10,000 images in which the red dot was at the center of the polygon with fewer sides ("Fewer" trials). Sample images are shown in Figure 8.

4.1.1.3. Training

As in Experiments 1 and 2, we trained AlexNet using both restricted and unrestricted transfer learning. We trained with 80% of our training data, withholding 20% as a validation set. All other training parameters were the same as in Experiment 1. Training concluded after 10 epochs or after the error rate on the validation set increased in six consecutive trials.

Training on ResNet-50 followed the same procedure as Experiment 2.

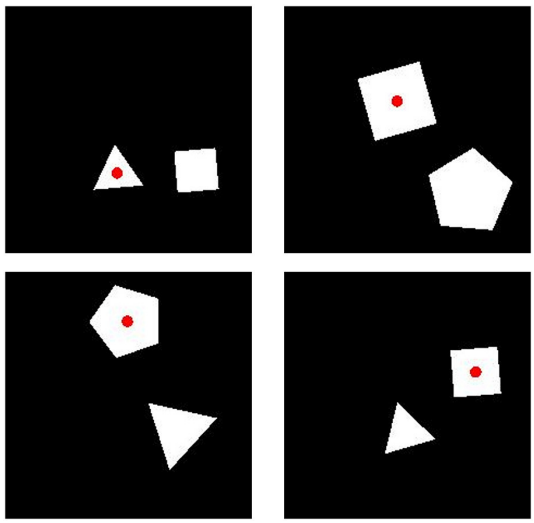

FIGURE 8
Sample training images for Experiment 3a. **(Top)** Two "Fewer" images. **(Bottom)** Two "More" images. Category membership was determined by the position of the small, red dot, placed on the polygon with either more or fewer sides.

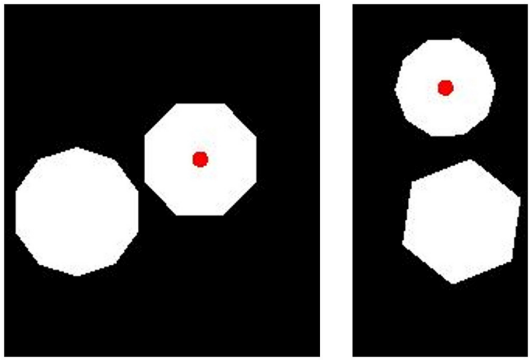

FIGURE 9
Sample images for the generalization test in Experiment 3b. **(Left)** A "Fewer" image. **(Right)** A "More" image. Category membership was determined by the position of the small, red dot, placed on the polygon with either more or fewer sides.

4.1.2. Results

Under restricted transfer learning, AlexNet trained to criterion after three epochs and achieved a classification accuracy of 84.4% on the validation set. Under unrestricted transfer learning, AlexNet took eight epochs to train to criterion and achieved a classification accuracy of 99.7% on the validation set, whereas ResNet-50 took 10 epochs to train to a final classification accuracy of 100%.

4.2. Experiment 3b: Generalization to other polygons

Despite testing a higher-order relation, network training in both restricted and unrestricted transfer learning was more successful than in either of our previous experiments. The crucial question, however, is whether the network learned response labels for particular concrete features of displays or whether the networks learned the abstract relation between dot location and the relative number of sides of a polygon. In Experiment 3b, we tested this question by generating new test images with polygons with more sides than those to which the network was exposed during training.

4.2.1. Method

In our generalization test, we created images with pairs of polygons that had twice as many sides as those present in training images. We replaced all three-sided polygons with six-sided polygons, all four-sided polygons with eight-sided polygons, and all five-sided polygons with ten-sided polygons. In all other respects, the test images were identical to the training images. We produced 50 "More" images in which the dot was placed on the polygon with more sides and 50 "Fewer" images in which the dot was placed on the polygon with fewer sides. Sample test images are shown in Figure 9.

Because AlexNet trained with restricted transfer learning also reached above-chance responding on the validation set, we tested it on the generalization task as well as both networks trained with unrestricted transfer learning.

4.2.2. Results

AlexNet trained with restricted and unrestricted transfer learning had an accuracy of 51% and 50% respectively on the generalization task. ResNet-50 trained with unrestricted transfer learning also had an accuracy of 50% on the generalization task. When we looked into how the networks were responding we found that the network trained with unrestricted transfer learning classified all of the "More" images correctly, but incorrectly classified all of the "Fewer" images as "More." The network trained with restricted transfer learning did the same apart from classifying one of the 50 "Fewer" images correctly.

4.3. Experiment 3c: Comparison with humans

While DCNNs appear able to learn to do the Experiment 3 task in a narrow sense, they showed no generalization whatsoever to other shapes. Performance in the generalization test was even worse for Experiment 3 than Experiments 1 or 2. One reason might be that in Experiment 3, we tested a higher-order perceptual relation than in previous experiments. In Experiment 3c, we tested humans on the same task to see if humans are capable of learning the more abstract relation between stimulus features required for accurate responding in Experiments 3a and 3b.

4.3.1. Method

4.3.1.1. Participants

Twelve participants (seven female, five male, $M_{age} = 21.0$) participated in Experiment 3c. Eight participants were recruited

from Loyola University and completed the experiment for course credit and four others were recruited from the University of California, Los Angeles and completed the experiment as volunteers. All participants were naive to the purpose of the experiment before participating.

4.3.1.2. Design

Experiment 3c consisted of a learning phase with 150 trials and a generalization phase with 40 trials.

4.3.1.3. Stimuli

Stimuli from the learning phase were randomly chosen from the network training data (Experiment 3a). Stimuli from the generalization phase were randomly chosen from the network generalization test (Experiment 3b).

4.3.1.4. Procedure

During the learning phase, images were presented in the center of the screen and participants were instructed to classify them into two arbitrary categories ("Category 1" or "Category 2") with no prior instruction on how to categorize images. Participants were given feedback after each trial and were told to try to discover the correct way of classifying images.

The generalization phase was the same as the learning phase except participants did not receive feedback after they responded.

4.3.2. Results

The results of Experiment 3c are shown in Figure 10. We found no significant difference between the first block of the training phase and either of the two subsequent blocks [$t_{(12)} < 1.91$, $p > 0.08$]. Participants performed significantly better than chance in all three training blocks [1st block: $t_{(12)} = 4.89$, $p < 0.001$; 2nd block: $t_{(12)} = 5.91$, $p < 0.001$; 3rd block: $t_{(12)} = 5.02$, $p < 0.001$].

As in Experiments 1 and 2, participants' learning during the training phase generalized when tested with polygons with more sides. Participants performed significantly better than chance in the generalization task, $t_{(12)} = 4.59$, $p < 0.001$. Performance on the generalization task did not significantly differ from performance on the third block of training, $t_{(12)} = 0.72$, $p = 0.48$.

4.4. Experiment 3a–c discussion

As in Experiments 1a and 2a, the networks learned to classify following unrestricted training. AlexNet also learned to classify well above chance performance following restricted transfer learning. Success in the restricted transfer learning case suggests that the features necessary for correct classification of ImageNet exemplars could be repurposed for the current classification task.

Still, neither restricted nor unrestricted transfer learning generalized to a different set of polygons that could be classified by the same rule. Specifically, the network failed to correctly classify polygons with twice as many sides as the training set. Once again, the data indicate that the network did not learn to classify based on a relational property that would generalize to other objects.

The performance of the networks in this study, and to some extent in the earlier studies, raises the interesting question of what was learned by the DCNNs? This is both theoretically interesting in its own right as well as relevant to distinguishing performance that arises from relational encoding from other variables in training displays that may allow powerful networks to exhibit behavior that could naively be interpreted as evidence of relational encoding. In general, it is hard to determine what properties DCNNs use in their responses. Neural networks in general may be characterized as carrying their knowledge in connection weights rather than in explicitly encoded properties. Moreover, the size of contemporary DCNNs allows for a vast array of stimulus variables to influence responses, and even with probing of node responses at various layers, there is no requirement that the properties captured in the network will be intelligible to humans. With regard to the present results, we consider one speculative hypothesis that illustrates how the network achieved some success in training without capturing the abstract relationship in the experiment-defined categorization task. Consider first that the network did learn to classify the polygons in the training set successfully, even following restricted transfer learning, and without, apparently, developing any explicit sensitivity to each polygon's number of sides. If so, it could be that a local feature that distinguishes the polygons, e.g., the internal angles of its vertices, was used in part for the classification task. Sensitivity to a feature like this would not be particularly surprising, given that the ImageNet training set contains many classes of artifacts, including rigid objects, for which the presence of vertices with specific angles might aid in identification. Prior research suggests that DCNNs adeptly capture local shape features (e.g., Baker et al., 2018).

In the initial training, with regular triangles, squares, and pentagons, when the probe was close to a vertex with angle = 108° (regular pentagon), the answer was "yes" (more). When the probe was close to a vertex with angle = 60° (equilateral triangle), the answer was "no" (fewer). A small set of slightly more complicated conjunctive rules allows for classification of the remaining cases without explicitly encoding the relation more-fewer sides. This learning would not, of course, generalize to a different set of polygons with different internal angles.

We expected the more-fewer relationship to be salient to human participants, leading to quick learning and full generalization. This appeared to be the case for the majority of our participants, who classified with >85% accuracy by the end of training and in generalization to polygons with more sides. However, with the added complexity of this classification, relative to Experiments 1 and 2, some participants may have found the perceptual more-fewer judgment too challenging or applied an idiosyncratic strategy. For example, one participant had high performance in training but showed little generalization, a pattern of behavior consistent with learning a complicated conjunctive rule (e.g., red dot in square + triangle = category A, red dot in square + pentagon = category B, etc.) that would have no utility for the different shapes. Another participant had performance in training and generalization testing well above chance, but below the level that would be expected had the more-fewer rule been learned. This participant may have been attempting to classify based on more-fewer sides, but never achieved high performance either because the task was too difficult for them, or perhaps due to poor attention or effort.

5. General discussion

The ability to extract abstract visual relations is crucial to many of the most important perceptual processes in human vision,

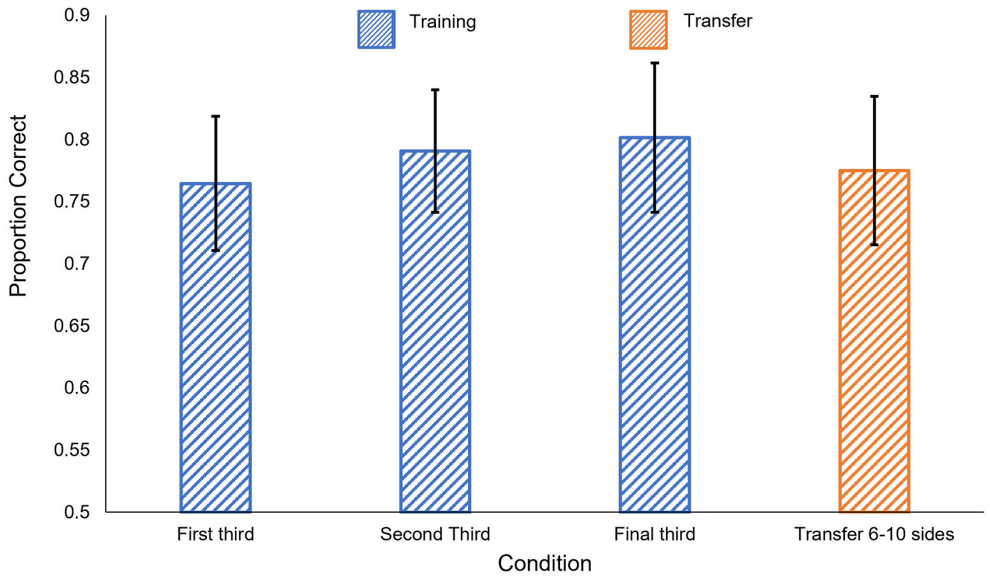

FIGURE 10
Human results in Experiment 3c. Blue: performance in the training phase, separated into 50-trial blocks. Orange: performance on the generalization tests. Error bars show ± one standard error of the mean.

including encoding of shape, arrangement, and structure in scenes, and perception of meaningful properties, such as animacy and causality in events. The notion of abstraction has a range of possible meanings (see Barsalou, 2003, for a useful discussion), but here, we intend a logical sense in which an abstract visual relation is one that involves a predicate that can be detected or represented despite having variable arguments. In perception, this idea is implicit in J.J. Gibson's theorizing about the role of "higher-order variables" in perception (e.g., Gibson, 1979), and more contemporary accounts of abstraction in perception and cognition have emphasized this notion (Marcus, 2001; Hummel, 2011; Kellman and Massey, 2013; Baker et al., 2020a). For present purposes, the impact is that detecting and utilizing abstract stimulus properties requires representations in which the argument is distinct from the relation. For example, a cluster of black pixels in between two clusters of white pixels is a relation, but not necessarily an abstract relation. An alternating ABA pattern of pixels irrespective of the pixel values would be an example of an abstract relation. While deep convolutional neural networks can evolve sensitivity to a vast array of possible "concrete" relations, and these no doubt underwrite their high classification accuracy in particular tasks, it is not clear that they have any access to abstract relations.

In three experiments, we tested DCNNs' ability to learn three abstract visual relations: same-different, inside-outside, and more-fewer. These certainly do not constitute an exhaustive test for all abstract relations, but there are reasons to believe they give valuable insight into DCNNs' general capability of learning abstract relations.

First, each of the three relations we tested depends on a different set of stimulus properties. Same-different depends on the comparison of contours across scale and position, inside/outside depends on the relative positions of the probe dot and a closed contour, and more-fewer depends on the comparison of magnitudes–either a polygon's number of sides or the angular size of its corners. A deficiency in processing any one of these stimulus features might account for insensitivity to one particular abstract relation, but a deficiency in all three relations points to a more general insensitivity to relations of an abstract nature.

Second, the three relations we tested are generally simple and are arguably relevant to systems that use visual information to extract ecologically relevant information from scenes. Experiments 1 and 2 tested what we call first-order relations, or relations between two image properties. Experiment 3 tested a second-order relation between first-order relations. All three are likely to be handled perceptually, given our brief exposure durations and rapid acquisition by most participants from classification feedback alone, and perception of relations in these cases is consistent with other research indicating the perceptual pickup of meaningful relations in scenes and events (Kellman and Massey, 2013; Hafri and Firestone, 2021). These relations all pick up on image features that could be important for object recognition, the task these networks were originally trained to perform. It is therefore reasonable to ask whether relations involving them can be learned in ImageNet-trained DCNNs. These are also the sorts of relations that may be useful in a variety of contexts where meaningful descriptions of objects, spatial layout, and events are to be acquired through visual perception.

The extraction of abstract relations as described here may account for discrepancies previously reported between successful DCNNs and human processing of objects and shape. In human vision, global shape is an abstract encoding in which relations are encoded but the particular sensory elements that act as carriers for relations are often transient, not surviving into more durable representations of objects and shape (Baker and Kellman, 2018). That shape is an abstract, configural notion accounts for the effortless recognition of similarity of shape despite changes in size, orientation, or constituent sensory elements. For example, a relatively small number of rectangles can make an easily recognized giraffe provided that their relative sizes and orientations are appropriate. Even for simple novel shapes, the abstract relations between elements are more important than

physical properties of the elements (Baker and Kellman, 2018). The observed incapacity of DCNNs to classify objects based on global shape information likely relates to the general absence of mechanisms that can capture and generalize abstract relations.

We used two training paradigms to assess apprehension of abstract visual relations. In restricted transfer learning, only the weights between the last representational layer and the decision layer were modified by training on a new classification task. In Experiments 1 and 2, we found no improvement in DCNN classification after a full 10 epochs using restricted transfer learning. This suggests that no weighted combination of features learned in ImageNet training could discriminate shapes based on sameness or enclosure. In Experiment 3, AlexNet reached above-chance classification accuracy with restricted transfer learning, indicating that certain features in the image are detected using learned filters from ImageNet training and can be used to discriminate between polygons with more sides and polygons with fewer sides, at least up to 84% accuracy and as long as shapes are within the distribution of polygons on which the network is trained. One possibility is that the network is already sensitive to local features like the angle of corners which can then be associated with distance from the probe dot.

We also tested both AlexNet and ResNet-50 using unrestricted transfer learning, in which all connection weights can be updated. In unrestricted transfer learning, DCNNs can learn new features that might be useful for a specific classification task. In all but one case, unrestricted transfer learning allowed DCNNs to reach performance levels significantly better than chance on the training task itself; however, in the unrestricted transfer learning for Experiment 1, ResNet-50 did not achieve above-chance performance even on the training data.

Most crucial for the questions motivating the present work was whether the networks had achieved training performance in each case by extraction of abstract visual relations or by some other rule that might not be intuitive to humans. We tested this by generating new testing stimuli whose individual features differed from those upon which the networks were originally trained, but could still be classified by the same abstract visual relation. If the abstract relation had been learned, then the network should have classified the new stimuli at the same level of accuracy it had reached on the training data.

Instead, we found that both networks' performance fell off substantially–often to around chance levels–when presented with new stimuli in which the same relations, if detected and used, would have produced perfect performance. The networks' lack of generalization strongly suggests that their improved performance on the training data was due to learning to classify based on a set of stimulus features that were specific to the kinds of images used during training (see Puebla and Bowers, 2021, for convergent evidence). For example, in Experiment 3, they may have learned some conjunctive rule about the kinds of polygons used in training rather than a rule about more or fewer sides that was divorced from the relation's arguments.

The lack of use of abstract visual relations was demonstrated particularly starkly in Experiment 2, where we placed the probe dot at all points within a single image and analyzed the network's pattern of responses. The network's "Inside" responses appeared to depend very little on the features of nearby contours or other relational properties that are easily describable by humans.

This lack of generalization suggests that deep convolutional networks are unable to disentangle relations from the arguments that fill them. In other words, a network might learn to say "Same" when two squares are on the screen, or when two circles are on the screen, but it is doing so in a "conjunctive" manner (Hummel, 2011); the learned relation binds the concrete stimulus features to the response, such that the network will not automatically generalize to say "Same" when two triangles are on the screen. Separating fillers from relations might require symbolic computation, something that does not appear to emerge spontaneously in the training of DCNNs.

We tested human participants with all of the relations presented to DCNNs. In contrast to the networks, humans easily learned all three of the abstract visual relations, often achieving ceiling performance levels in the first 50 training examples. More importantly, human performance was robust in generalization tests with stimuli having features different from than the training data. Across all three experiments, we found no significant difference between human performance on any of the generalization tasks and the last 50 trials in which they were training with feedback.

This difference between humans and networks points to humans' remarkable ability to perceive and use abstract visual relations. It has been argued that even what appear to be simple, basic visual tasks in human visual perception involve abstraction (Kellman and Massey, 2013; Baker and Kellman, 2018). The results presented here show that there are alternative intelligent systems that can be very successful at similar tasks (e.g., image classification) without human-like sensitivity to abstract relations.

Differences between humans and DCNNs also provide a striking example of the flexibility of human visual perception in contrast with the relative inflexibility of processing in deep network architectures. Whereas, humans were able to learn new visual tasks within a few dozen trials of initial exposure, even after tens of thousands of trials, DCNNs were incapable of learning them. Humans' superior flexibility is in one sense unsurprising because, unlike DCNNs, humans are adapted to perform a variety of visual routines that goes far beyond image classification. On the other hand, the case of abstract visual relations is interesting because encoding relations abstractly might crucially underpin our more general flexibility. For example, consider the enclosure relation we examined in Experiment 2. Knowing whether a visual feature is intrinsic to an object or merely correlates with the object can be partly determined by whether it is enclosed by the object's bounding contour. Binding features to objects furnishes a great deal of flexibility in learning about new objects, but it is hard to see how this flexibility and transfer can be accomplished without some representation of abstract notions such as object, boundary, figure vs. ground, etc. Other work suggests that DCNNs do not naturally acquire such representations, such as segmenting the image into figure and ground when learning to classify novel objects (Baker et al., 2018, 2020b).

From the perspective of deep networks, an inability to learn abstract visual relations might be predictive of poor performance on a wide array of visual routines. Processes like segmenting figure from ground (Peterson and Salvagio, 2008), completing an object behind an occluder (Kellman and Shipley, 1991), judging the causality of an event (Michotte, 1954), and representing the shape of objects (Koffka, 1935; Kubovy and Wagemans, 1995; Baker and Kellman, 2018) all depend on access to abstract relations in human vision.

DCNNs may be able to learn appropriate responses in a training set of displays, but without the ability to learn abstract relations, they will perform them in a very different way from humans. An example

of this can be seen in comparisons between human and DCNN shape sensitivity. DCNNs do use some shape information (although to a lesser extent than humans), but they use different aspects of shape from humans (Baker et al., 2018, 2020b). These differences can lead to surprising errors in DCNNs, as when an adversarial attack that would be unnoticeable to humans completely changes a network's classification (Szegedy et al., 2013). In the same way, DCNNs might be able to learn responses to other important visual tasks, but without the use of relations. Consequently, we expect that DCNN learning will in general be less robust, and vulnerable to errors that humans would be unlikely to expect (and therefore, in high stakes domains, potentially much more hazardous).

How might DCNNs be enhanced to retain their valuable abilities to learn visual classifications but to also capture abstract visual relations? This is a difficult question to answer because the convolution operators underpinning DCNN operations may be ill suited for the task. Recent ImageNet-trained recurrent (Kubilius et al., 2019) and attention-based (Dosovitskiy et al., 2020) architectures have shown better and more humanlike performance on several tasks, but do not appear to be more sensitive to the global shape of objects (Baker and Elder, 2022). It remains unknown whether a new architecture paired with training data more targeted toward apprehension of visual relations would produce the kind of abstraction observed in humans.

In our view, a more extreme adjustment to these networks might be needed. As argued by Hummel (2011), abstract visual relations might require symbolic processing to separate roles from their fillers. Animal studies have shown that many animals fail to complete same-different tasks that depend on abstract relations (Gentner et al., 2021). However, chimpanzees that are exposed to training with symbolic systems are able to perform well on same-different tasks that chimpanzees with non-symbolic training can not do (Premack, 1983).

Research into symbolic networks has demonstrated that they can represent the spatial relations between parts to build up structural descriptions (Hummel and Stankiewicz, 1996; Hummel, 2001) and to generalize to novel instances of shapes based on their relations (Kellman et al., 1999). It remains unclear how to combine symbolic processing with deep convolutional networks. Some related work on large artificial networks in linguistics (e.g., Vankov and Bowers, 2020; Jiang et al., 2021; Kim and Smolensky, 2021) suggests some strategies for combining extensive associative training with symbolic processing. In vision, capsule networks (Sabour et al., 2018) include some relational coding and have been shown to increase configural sensitivity in uncrowding effects (Doerig et al., 2020). Another recent model adds external memory to a recurrent DCNN to allow for explicit symbolic processing, resulting in rapid abstract rule learning (Webb et al., 2021).

6. Conclusion

DCNNs are remarkably accurate image classifiers that, to some degree, mimic human behavior and neurophysiology. These similarities, however, distract from the fact that DCNNs learn very different kinds of visual relations than humans. While humans readily learn relations separable from their arguments, we found no evidence that arguments and their relations are separable in DCNNs. This difference is of fundamental importance. While DCNNs have access to non-abstract relational encoding sufficient for, e.g., human-like performance levels of object recognition, they lack a critical form of representation that supports more general visual perception and reasoning.

Any apparent visual reasoning performed by a conventional DCNN appears to rely on complex mappings among encodings of relatively concrete stimulus properties, rather than any abstract representation of visual information. We believe that this limitation will become more apparent as DCNNs are trained to perform a wider variety of human visual tasks, and may not be overcome with larger, more complex networks. Instead, alternative architectures, possibly ones that explicitly include symbolic computations, and/or modified training regimes, will be needed for DCNNs to apprehend abstract visual relations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by IRB, Loyola University Chicago and IRB, UCLA. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

NB: conceptualization, design, coding, testing participants, simulations, writing, and data analysis. PG: conceptualization, design, coding, and writing. AP: coding, testing participants, and editing. PK: conceptualization, design, and writing. All authors contributed to the article and approved the submitted version.

Funding

We gratefully acknowledge support from the National Institutes of Health award number: R01 CA236791 to PK.

Acknowledgments

Portions of this work were presented at the 2021 meeting of the Vision Sciences Society (VSS) and the 2021 meeting of the Configural Processing Consortium (CPC). We thank Hongjing Lu for helpful discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adelson, E. H., and Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature* 300, 523–525. doi: 10.1038/300523a0

Baker, N., and Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience* 2022, 104913. doi: 10.1016/j.isci.2022.104913

Baker, N., Garrigan, P., and Kellman, P. J. (2020a). Constant curvature segments as building blocks of 2D shape representation. *J. Exp. Psychol. Gen*. 2020, xge0001007. doi: 10.1037/xge0001007

Baker, N., and Kellman, P. J. (2018). Abstract shape representation in human visual perception. *J. Exp. Psychol. Gen*. 147, 1295. doi: 10.1037/xge0000409

Baker, N., and Kellman, P. J. (2021). Constant curvature modeling of abstract shape representation. *PLoS ONE* 16, e0254719. doi: 10.1371/journal.pone.0254719

Baker, N., and Kellman, P. J. (in press). Independent mechanisms for processing local contour features and global shape. *J. Exp. Psychol. Gen.*

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol*. 14, e1006613. doi: 10.1371/journal.pcbi.1006613

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2020b). Local features and global shape information in object classification by deep convolutional neural networks. *Vis. Res*. 172, 46–61. doi: 10.1016/j.visres.2020.04.003

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philos. Trans. Royal Soc. B. Biol. Sci*. 358, 1177–1187. doi: 10.1098/rstb.2003.1319

Bell, J., Badcock, D. R., Wilson, H., and Wilkinson, F. (2007). Detection of shape in radial frequencycontours: Independence of local and global form information. *Vis. Res*. 47, 1518–1522 doi: 10.1016/j.visres.2007.01.006

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pat. Anal. Machine Intell*. 24, 509–522. doi: 10.1109/34.993558

Bergevin, R., and Levine, M. D. (1993). Generic object recognition: Building and matching coarse descriptions from line drawings. *IEEE Trans. Pat. Anal. Machine Intell*. 15, 19–36. doi: 10.1109/34.184772

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychol. Rev*. 94, 115. doi: 10.1037/0033-295X.94.2.115

Biederman, I., and Ju, G. (1988). Surface vs. edge-based determinants of visual recognition. *Cogn. Psychol*. 20, 38–64. doi: 10.1016/0010-0285(88)90024-2

Burr, D. C., Anobile, G., and Arrighi, R. (2017). Psychophysical evidence for the number sense. *Philos. Trans. Royal Soc. B Biol. Sci*. 373, 20170045. doi: 10.1098/rstb.2017.0045

Cheng, X., Lin, C., Lou, C., Zhang, W., Han, Y., Ding, X., et al. (2021). Small numerosity advantage for sequential enumeration on RSVP stimuli: An object individuation-based account. *Psychol. Res*. 85, 734–763. doi: 10.1007/s00426-019-01264-5

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL), 248–255. doi: 10.1109/CVPR.2009.5206848

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., and Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLoS Comput. Biol*. 16, e1008017. doi: 10.1371/journal.pcbi.1008017

Donderi, D. C., and Zelnicker, D. (1969). Parallel processing in visual same-different decisions. *Percept. Psychophys*. 5, 197–200. doi: 10.3758/BF03210537

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv [Preprint] arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929

Driver, J., and Baylis, G. C. (1996). Edge-assignment and figure–ground segmentation in short-term visual matching. *Cogn. Psychol*. 31, 248–306. doi: 10.1006/cogp.1996.0018

Dubey, R., Peterson, J., Khosla, A., Yang, M. H., and Ghanem, B. (2015). "What makes an object memorable?" in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago), 1089–1097. doi: 10.1109/ICCV.2015.130

Elder, J., and Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vis. Res*. 33, 981–991. doi: 10.1016/0042-6989(93)90080-G

Elder, J. H., and Velisavljević, L. (2009). Cue dynamics underlying rapid detection of animals in natural scenes. *J. Vis*. 9, 7–7. doi: 10.1167/9.7.7

Garrigan, P. (2012). The effect of contour closure on shape recognition. *Perception* 41, 221–235. doi: 10.1068/p7145

Garrigan, P., and Kellman, P. J. (2008). Perceptual learning depends on perceptual constancy. *Proc. Natl. Acad. Sci. U. S. A*. 105, 2248–2253. doi: 10.1073/pnas.0711878105

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F., and Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv [Preprint]*. arXiv: 1811.12231.

Gentner, D., Shao, R., Simms, N., and Hespos, S. (2021). Learning same and different relations: cross-species comparisons. *Curr. Opin. Behav. Sci*. 37, 84–89. doi: 10.1016/j.cobeha.2020.11.013

Gibson, E. J. (1969). *Principles of Perceptual Learning and Development*. New York, NY: Appleton-Century-Crofts.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. New York, NY: Houghton Mifflin.

Greff, K., Van Steenkiste, S., and Schmidhuber, J. (2020). On the binding problem in artificial neural networks. arXiv [Preprint] arXiv:2012.05208. doi: 10.48550/arXiv.2012.05208

Hafri, A., and Firestone, C. (2021). The perception of relations. *Trends Cogn. Sci*. 25, 475–492. doi: 10.1016/j.tics.2021.01.006

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CPVR)* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

He, L., Zhang, J., Zhou, T., and Chen, L. (2009). Connectedness affects dot numerosity judgment: Implications for configural processing. *Psychonom. Bull. Rev*. 16, 509–517. doi: 10.3758/PBR.16.3.509

Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *Am. J. Psychol*. 57, 243–259. doi: 10.2307/1416950

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Vis. Cogn*. 8, 489–517. doi: 10.1080/13506280143000214

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connect. Sci*. 23, 109–118. doi: 10.1080/09540091.2011.569880

Hummel, J. E., and Stankiewicz, B. J. (1996). "An architecture for rapid, hierarchical structural description," in *Attention and Performance XVI: Information Integration in Perception and Communication* (Cambridge, MA: MIT Press), 93–121.

Izard, V., Sann, C., Spelke, E. S., and Streri, A. (2009). Newborn infants perceive abstract numbers. *Proc. Natl. Acad. Sci. U. S. A*. 106, 10382–10385. doi: 10.1073/pnas.0812142106

Jacob, G., Pramod, R. T., Katti, H., and Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat. Commun*. 12, 1–14. doi: 10.1038/s41467-021-22078-3

Jiang, Y., Celikyilmaz, A., Smolensky, P., Soulos, P., Rao, S., Palangi, H., et al. (2021). Enriching transformers with structured tensor-product representations for abstractive summarization. arXiv [Preprint] arXiv:2106.01317. doi: 10.18653/v1/2021.naacl-main.381

Johansson, G. (1978). "Visual event perception," in *Perception*, eds R. Held, H. W. Leibowitz, and H. L. Teuber (Berlin, Heidelberg: Springer), 675–711. doi: 10.1007/978-3-642-46354-9_22

Kanizsa, G. (1979). *Organization in Vision: Essays on Gestalt Perception*. Westport, CT: Praeger Publishers.

Kellman, P. J., Burke, T., and Hummel, J. E. (1999). "Modeling perceptual learning of abstract invariants," in *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, eds M. Hahn and S. C. Stoness (Hillsdale, NJ: Erlbaum), 264–269. doi: 10.4324/9781410603494-51

Kellman, P. J., and Fuchser, V. (in press). "Visual completion and intermediate representations in object formation," in *Sensory Individuals: Contemporary Perspectives on Modality-specific and Multimodal Perceptual Objects*, eds A. Mroczko-Wasowicz and R. Grush (London: Oxford University Press).

Kellman, P. J., and Massey, C. M. (2013). "Perceptual learning, cognition, and expertise," in *The Psychology of Learning and Motivation, Vol. 58*, ed B. H. Ross (Amsterdam: Elsevier Inc), 117–165. doi: 10.1016/B978-0-12-407237-4.00004-9

Kellman, P. J., and Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cogn. Psychol*. 23, 141–221. doi: 10.1016/0010-0285(91)90009-D

Kim, J., Ricci, M., and Serre, T. (2018). Not-So-CLEVR: Learning same-different relations strains feedforward neural networks. *Interface Focus* 8, 20180011. doi: 10.1098/rsfs.2018.0011

Kim, N., and Smolensky, P. (2021). Testing for grammatical category abstraction in neural language models. *Proc. Soc. Comput. Linguist*. 4, 467–470. doi: 10.7275/2nb8-ag59

Kimchi, R. (1998). Uniform connectedness and grouping in the perceptual organization of hierarchical patterns. *J. Exp. Psychol.* 24, 1105. doi: 10.1037/0096-1523.24.4.1105

Klatzky, R. L., Wu, B., and Stetten, G. (2008). Spatial representations from perception and cognitive mediation: The case of ultrasound. *Curr. Direct. Psychol. Sci.* 17, 359–364. doi: 10.1111/j.1467-8721.2008.00606.x

Koffka, K. (1935). *Principles of Gestalt Psychology*. London: Routledge.

Kovacs, I., and Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proc. Natl. Acad. Sci. U. S. A.* 90, 7495–7497. doi: 10.1073/pnas.90.16.7495

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc), 1097–1105.

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* 12, e1004896. doi: 10.1371/journal.pcbi.1004896

Kubilius, J., Schrimpf, M., Hong, H., Kar, K., Majaj, N. J., Rajalingham, R., et al. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Adv. Neural Inform. Process. Syst.* 32, 6161. doi: 10.48550/arXiv.1909.06161

Kubovy, M., and Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: A quantitative Gestalt theory. *Psychol. Sci.* 6, 225–234. doi: 10.1111/j.1467-9280.1995.tb00597.x

Kümmerer, M., Theis, L., and Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv [Preprint] arXiv:1411.1045. doi: 10.48550/arXiv.1411.1045

Lloyd-Jones, T. J., and Luckhurst, L. (2002). Outline shape is a mediator of object recognition that is particularly important for living things. *Mem. Cogn.* 30, 489–498. doi: 10.3758/BF03194950

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proc. Seventh IEEE Int. Conf. Comput. Vis.* 2, 1150–1157. doi: 10.1109/ICCV.1999.790410

Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1187.001.0001

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Messina, N., Amato, G., Carrara, F., Gennaro, C., and Falchi, F. (2021). Solving the same-different task with convolutional neural networks. *Pat. Recogn. Lett.* 143, 75–80. doi: 10.1016/j.patrec.2020.12.019

Michotte, A. (1954). *The Perception of Causality*. London: Routledge.

Michotte, A., Thinès, G., and Crabbé, G. (1964). *Les complements amodaux des structures perceptives*. Louvain: Publications Universitaires.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cogn. Psychol.* 9, 353–383. doi: 10.1016/0010-0285(77)90012-3

Palmer, E. M., Kellman, P. J., and Shipley, T. F. (2006). A theory of dynamic occluded and illusory object perception. *J. Exp. Psychol. Gen.* 135, 513. doi: 10.1037/0096-3445.135.4.513

Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. arXiv [Preprint] arXiv:1608.02164. doi: 10.24963/ijcai.2017/697

Peterson, M. A., and Salvagio, E. (2008). Inhibitory competition in figure-ground perception: Context and convexity. *J. Vis.* 8, 1–13. doi: 10.1167/8.16.4

Piazza, M., Fumarola, A., Chinello, A., and Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition* 121, 147–153. doi: 10.1016/j.cognition.2011.05.007

Pizlo, Z. (2008). *3D Shape*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/7705.001.0001

Pospisil, D. A., Pasupathy, A., and Bair, W. (2018). "Artiphysiology" reveals V4-like shape tuning in a deep network trained for image classification. *Elife* 7, e38242. doi: 10.7554/eLife.38242

Premack, D. (1983). The codes of man and beasts. *Behav. Brain Sci.* 6, 125–136. doi: 10.1017/S0140525X00015077

Puebla, G., and Bowers, J. (2021). Can deep convolutional neural networks learn same-different relations? *Proc. Ann. Meet. Cogn. Sci. Soc.* 43, 8551. doi: 10.1101/2021.04.06.438551

Rezanejad, M., and Siddiqi, K. (2013). "Flux graphs for 2D shape analysis," in *Shape Perception in Human and Computer Vision*, eds S. Dickinson and Z. Pizlo (Berlin/?Heidelberg?: Springer), 41–54. doi: 10.1007/978-1-4471-5195-1_3

Rubin, E. (1915/1958). "Visuell wahrgenommene figuren (Copenhagen: Gyldenalske Boghandel, 1915); reprinted as Figure and ground," in *Readings in Perception*, ed D. C. Beardslee (Princeton, NJ: D. van Nostrand), 194–203.

Sabour, S., Frosst, N., and Hinton, G. (2018). "Matrix capsules with EM routing," in *6th International Conference on Learning Representations, ICLR* (Vancouver, BC), 115.

Scholl, B. J., and Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends Cogn. Sci.* 4, 299–309. doi: 10.1016/S1364-6613(00)01506-0

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pat. Anal. Machine Intell.* 22, 888–905. doi: 10.1109/34.868688

Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vis.* 81, 2–23. doi: 10.1007/s11263-007-0109-1

Stabinger, S., Rodríguez-Sánchez, A., and Piater, J. (2016). "25 years of cnns: Can we compare to human abstraction capabilities?" in *Artificial Neural Networks and Machine Learning – ICANN 2016*, eds A. Villa, P. Masulli, A. Pons Rivero (Berlin/Heidelberg: Springer), 380–387. doi: 10.1007/978-3-319-44781-0_45

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. arXiv [Preprint] arXiv:1312.6199. doi: 10.48550/arXiv.1312.6199

Ullman, S. (1979). The interpretation of structure from motion. *Proc. Royal Soc. Lond. Ser. B Biol. Sci.* 203, 405–426. doi: 10.1098/rspb.1979.0006

Vankov, I. I., and Bowers, J. S. (2020). Training neural networks to encode symbols enables combinatorial generalization. *Philos. Trans. Royal Soc. B.* 375, 20190309. doi: 10.1098/rstb.2019.0309

Wallach, H., and O'Connell, D. N. (1953). The kinetic depth effect. *J. Exp. Psychol.* 45, 205. doi: 10.1037/h0056880

Webb, T. W., Sinha, I., and Cohen, J. D. (2021). Emergent symbols through binding in external memory. arXiv [Preprint] arXiv:2012.14601. doi: 10.48550/arXiv.2012.14601

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Zhou, H., Friedman, H. S., and Von Der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611. doi: 10.1523/JNEUROSCI.20-17-06594.2000

Frontiers in
Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact