DAGSTUHL
MANIFESTOS

Volume 11, Issue 1, January – December 2025

ISSN 2193-2433

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at https://www.dagstuhl.de/dagpub/2193-2433

Publication date
December, 2025

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `https://dnb.d-nb.de`.

Aims and Scope
The manifestos from Dagstuhl Perspectives Workshops are published in the *Dagstuhl Manifestos* journal. Each manifesto aims for describing the state-of-the-art in a field along with its shortcomings, strengths. Based on this, position statements, perspectives for the future are illustrated. A manifesto typically has a less technical character; instead it provides guidelines, roadmaps for a sustainable organisation of future progress.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Publishing
Oktavie-Allee, 66687 Wadern, Germany
publishing@dagstuhl.de

Manifesto from Dagstuhl Perspectives Workshop 25122

Climate Change: What is Computing's Responsibility?

Bran Knowles[*1], Vicki L. Hanson[*2], Christoph Becker[†3], Mike Berners-Lee[†4], Andrew A. Chien[†5], Benoit Combemale[†6], Vlad Coroamă[†7], Koen De Bosschere[†8], Yi Ding[9], Adrian Friday[†10], Boris Gamazaychikov[†11], Lynda Hardman[†12], Simon Hinterholzer[†13], Mattias Höjer[†14], Lynn Kaack[†15], Lenneke Kuijer[†16], Anne-Laure Ligozat[†17], Jan Tobias Muehlberg[†18], Yunmook Nah[†19], Thomas Olsson[†20], Anne-Cécile Orgerie[†21], Daniel Pargman[†22], Birgit Penzenstadler[†23], Tom Romanoff[†24], Emma Strubell[†25], Colin Venters[†26], and Junhua Zhao[†27]

1 **Lancaster University, Lancaster, UK.** b.h.knowles1@lancaster.ac.uk
2 **Association for Computing Machinery, New York, USA.** vicki.hanson@hq.acm.org
3 **University of Toronto, CA**
4 **Lancaster University, GB**
5 **University of Chicago, US**
6 **University of Rennes, FR**
7 **Roegen Centre for Sustainability – Zürich, CH**
8 **Ghent University, BE**
9 **Purdue University – West Lafayette, US**
10 **Lancaster University, GB**
11 **Salesforce – Paris, FR**
12 **CWI – Amsterdam, NL; Utrecht University, NL**
13 **Borderstep Institute – Berlin, DE**
14 **KTH Royal Institute of Technology – Stockholm, SE**
15 **Hertie School of Governance – Berlin, DE**
16 **TU Eindhoven, NL**
17 **CNRS – Orsay, FR**
18 **Free University of Brussels, BE**
19 **Dankook University – Yongin-si, KR**
20 **University of Tampere, FI**
21 **CNRS – IRISA – Rennes, FR**
22 **KTH Royal Institute of Technology – Stockholm, SE**
23 **Chalmers University of Technology – Göteborg, SE**
24 **ACM – New York, US**
25 **Carnegie Mellon University – Pittsburgh, US**
26 **University of Limerick, IE**
27 **The Chinese University of Hong Kong – Shenzhen, CN**

Abstract

This Manifesto was produced from the Perspectives Workshop 25122 entitled "Climate Change: What is Computing's Responsibility?" held March 16–19, 2025 at Schloss Dagstuhl, Germany. The Workshop provided a forum for world-leading computer scientists and expert consultants on environmental policy and sustainable transition to engage in a critical and urgent conversation about computing's responsibilities in addressing climate change – or more aptly, climate crisis.[1]

[*] Editor / Organizer
[†] Contributor
[1] We adopt the singular form over "climate crises" for readability, but note the inextricability of human induced climate change from the primary and secondary effects of this change on all kinds of local and global essential systems which are also in a state of crisis.

Climate Change: What is Computing's Responsibility?, *Dagstuhl Manifestos*, Vol. 11, Issue 1, pp. 1–18
Editors: Bran Knowles and Vicki L. Hanson

Dagstuhl Manifestos
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The resulting Manifesto outlines commitments and directions for future action which, if adopted as a basis for more responsible computing practices, will help ensure that these technologies do not threaten the long-term habitability of the planet.

We preface our Manifesto with a recognition that humanity is on a path *that is not in agreement* with international global warming targets[2] [2] and explore how computing technologies are currently hastening the overshoot of these boundaries. We critically assess the vaunted potential for harnessing computing technologies for the mitigation of global warming, agreeing that, under current circumstances, computing is contributing to negative environmental impacts in other sectors. Computing primarily improves efficiency and reduces costs which leads to more consumption and more negative environmental impact. Relying solely on efficiency gains in computing has thus far proven to be insufficient to curb global greenhouse gas emissions. Therefore, computing's purpose within a strategy for tackling climate change must be reimagined.

Our recommendations cover changes that need to be urgently made to the design priorities of computing technologies, but also speak to the more systemic shift in mindset, with sustainability and human rights providing a necessary moral foundation for developing the kinds of computing technologies most needed by society. We also stress the importance of digital policy that accounts for both the direct material impacts of computing and the detrimental indirect impacts arising from computing-enabled efficiencies, and the role of computing professionals in informing policy making.

Perspectives Workshop March 16–19, 2025 – https://www.dagstuhl.de/25122

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Human-centered computing → Human computer interaction (HCI); Software and its engineering

Keywords and phrases sustainability, climate change, efficiency, supply chain management, climate modelling

Digital Object Identifier 10.4230/DagMan.11.1.1

Funding This workshop was initiated and partly funded by the ACM Europe Council [4], which promotes dialogue and the exchange of ideas on technology and computing policy issues with the European Commission and other governmental bodies in Europe.

Executive Summary

The increasing threat of climate crisis requires concrete efforts to minimise the negative impacts of computing practices today on current and future generations of life on Earth. While computing has enabled transformative progress, it is implicated in unsustainability, contributing to and accelerating the rapid overshoot of planetary boundaries [1, 59, 19, 62]. It is our responsibility to ensure that technological advances benefit human and non-human well-being and do not come at the cost of a just ecological foundation *for all life on Earth.*

Knowing that a multitude of planetary boundaries are already being crossed today [1, 59] and that digital technologies cause significant social impacts, this Manifesto focuses on impacts that exacerbate the ecological and climate crises. We are aware that social impacts are also important and in many cases, if conflicting, are typically valued higher. Ultimately, however, human well-being is contingent on being within a healthy environment and living within the limits of the planet that can sustain life.

[2] While we focus on climate change, we also note incompatibility with other planetary boundaries and UN sustainable development goals.

As a basis for our commitments, we recognise:

1. The need for a safe and just space for all life on the planet to thrive [58].
2. The urgent need to address environmental challenges, including the short time horizon for reducing greenhouse gas emissions.
3. The limited resources of the planet and the need to uphold planetary boundaries; and for those boundaries that have been surpassed, the need to move back within a safe region [59].
4. That, as with other resources, computational resources are not equitably distributed today. While some use excessive amounts of computational resources, others do not have enough.
5. The need to attend to environmental challenges while advancing universal human rights and well-being.
6. That computing is not immaterial and has both direct and indirect adverse impacts on the environment:
   a. Direct impacts occur through the energy used for, e.g., mineral extraction, the manufacturing, use and disposal of hardware that provides computation, storage, network access, and interactivity; but also include water consumption and ecological disruption at these stages.
   b. Indirect impacts arise from the transformative nature of computing, with effects across societal and economic sectors yielding both beneficial and detrimental environmental impacts.
7. That computing, under current circumstances, contributes to the negative environmental impacts of other sectors and is accelerating the exceeding of planetary boundaries. Humanity is on a path that is not in agreement with the 1.5-2°C of global warming target of the Paris Agreement [2], and computing is likely getting us there faster.
8. That the enablement narrative of direct impacts being outweighed by positive indirect impacts (sometimes called 'digital exceptionalism' [39]) is ignoring the detrimental indirect effects, cherry-picking for those domains or applications with a likely beneficial outcome, and thus likely not true.
9. The tendency for low-level interventions to induce rebound effects and confirm an unsustainable status-quo.
10. The need for changing high-level leverage points (for example societal structures and paradigms) even though it is difficult.
11. That limiting climate change is a systemic challenge. Advancing computing, and its uses, cognizant of environmental limits and impacts is the joint responsibility of governments, corporations, educational institutions and computing professionals worldwide.
12. The contested nature of sustainability [24] and prevalence of misunderstandings of the term, along with the principles of sustainability design outlined in the Karlskrona Manifesto [12].

We commit to:

1. Reducing the environmental impacts of computing systems, comprising direct, indirect, and systemic impacts.
2. Contributing to achieving absolute reductions in global greenhouse gas emissions.
3. Achieving absolute reductions in global greenhouse gas emissions caused by computing by:
   a. Ensuring durability, longevity, repairability, and reusability of hardware.
   b. Ensuring adaptability, longevity and resilience of software [32].
   c. Advocating sufficiency in use of computational resources.

   d. Aiming for Paris-compliant [2] reduction of around 7% p.a. of CO_2 emissions in the organisation and participation of Computer Science scientific and industry-oriented events (conferences, conventions, etc.).
4. Sharing information regarding potential and actual impacts of computing in a transparent, evidence-based, truthful, and holistic manner.
5. Promoting said information towards the media and general public.
6. Elevating sustainability to a first-class consideration for computing and computing facility design and implementation decisions.
7. Educating for climate-conscious computing with a sustainability mindset.
8. Leveraging computing to solve environmental challenges for climate solutions when relevant and advocating for policy/funding/innovation frameworks which ensure their effective uptake and usage while minimising rebound effects.
9. To this end, exploring, and where possible measuring, environmental and societal impacts; contributing to appropriate measurement methods and practices.
10. Seeking to effect positive change by locating and acting on leverage points that are beyond the conventional scope of computing.
11. Prioritising purposes that support sustainable development goals [50], well-being, and scientific truths.
12. Working to support the equitable and democratic redistribution of computing capabilities.
13. Advocating for national and international digital policy that focuses on climate and environmental sustainability.
14. Including all populations that are negatively affected by environmental impacts in these conversations.
15. All of the above while promoting and advancing human rights as defined by the Universal Declaration of Human Rights [7].

Table of Contents

1 Introduction

Global average temperatures are now about 1.1°C above pre-industrial levels, with most recent months consistently ranking among the hottest on record. At the same time, atmospheric CO_2 has surged past 420 ppm – the highest concentration in at least two million years – driving a relentless rise in greenhouse forcing. Under current national pledges, the world is headed for roughly 2.7°C of warming by 2100 (according to the Climate Action Tracker compiled by Germany-based nonprofits Climate Analytics and the NewClimate Institute[3]), far beyond the Paris Agreement's 1.5°C ambition and exposing a wide gap between commitments and what's needed. The IPCC warns that keeping warming under 1.5°C requires achieving net-zero CO_2 emissions by around 2050 and slashing methane and other short-lived pollutants – meaning "deep, rapid, and sustained" cuts this decade. Yet the latest UN global stocktake finds existing plans are woefully inadequate, hampered by financial shortfalls, uneven policy rollout, and slow deployment of renewables and efficiency measures. Without immediate, sweeping transformation across energy, land use, transport, and industry, irreversible impacts – accelerating sea-level rise, ecosystem collapse, and more frequent extreme weather – will become unavoidable, underscoring how narrow our window has become.

2 Positionality statement

This Manifesto was composed by computing professionals, researchers, educators, developers, and users from the Global North. We write from our limited perspective, recognising our comparative privilege in shaping possibilities and, with that, our outsized responsibility to shape these possibilities in accordance with sustainability and global justice.

We further recognise that many of the populations that are most negatively affected by environmental impacts of climate change are not included in the creation of this Manifesto (see "asymmetric vulnerability" [34]). We strongly recommend including *all* affected populations in these conversations – or, where direct participation is not possible, engaging methodologies that facilitate perspective taking [27]. This more inclusive approach is consistent with the commitment to social justice that this Manifesto underscores.

3 Computing and Climate Change

For an issue that has been publicly debated since the 1960s, concerns about computing's contribution to climate change (now crisis) have emerged surprisingly recently [40], with "Sustainable Human-Computer Interaction" [46, 14] and "Green IT" [49] entering the computing landscape around 2007. In the meantime, the total carbon footprint of computing has grown inexorably. At present, although the share of Information and Communication Technologies in global CO_2 emissions seems small at approximately 3% [33], collectively it exceeds that of many whole nations and is similar in magnitude to aviation, which is under significant pressure to reduce in line with internationally agreed climate targets. Moreover, mainly due to the rise of Generative AI, computing technologies are contributing to a rapid acceleration in global energy demand [37].

[3] `https://climateactiontracker.org/`

The increasing threat of the climate crisis requires concrete efforts to minimise the negative impacts of computing practices today on current and future generations of life on Earth.

3.1 Challenging optimistic rhetoric

While there are abundant examples of computing that benefits society, even "positive" applications can have attendant harms. Given their inherent accelerating properties (cf. amplification theories [64]), computing technologies are notorious for intensifying societal and economic trends. The year 2024 was the first to have an average temperature exceeding 1.5°C above pre-industrial levels [20], the preferred limit laid out in the historic Paris Agreement [2], and we are on a path to exceed the even riskier 2°C of global warming target [52]. Under business as usual, per amplification, computing is likely getting us past 2°C warming faster.

The improved efficiency that computing applications enable is typically viewed as naturally yielding reductions in energy use. In reality, efficiency can drive (and in our current paradigm, does drive) the amplification of social and environmental harms. The fact that greenhouse gas (GHG) emissions have been steadily rising over the decades that computing technologies have been delivering continuous efficiency improvements is evidence of the tendency for computing-enabled efficiency gains to generate increases in demand greater than the energy reductions derived from the efficiency (cf. Jevons' Paradox [5]). All available evidence indicates that computing contributes to negative environmental impacts of other sectors by making it cheaper and easier to do more.

Artificial Intelligence is often touted as delivering other critical environmental benefits, such as energy optimisation, materials discovery, resource planning, forecasting and disaster mitigation, policy analysis and modeling, and more. Many of these benefits have yet to be demonstrated in practice; meanwhile, the negative environmental impacts of AI are already quite clear and are significantly greater than these proposed benefits. The vast majority of AI applications are not even purported to benefit the environment, and much of the consumption we would characterise as wasteful, excessive, and even harmful. The selective highlighting of climate mitigation applications of AI represents strategic – and dangerously misleading – obfuscation of the already alarming and still growing environmental impacts of these technologies.

This Manifesto resoundingly rejects "digital exceptionalism" [39], i.e. the justification for unchecked growth on the grounds that direct negative impacts of computing are outweighed by their positive indirect impacts. This argument ignores the detrimental indirect effects, cherry-picking for those domains or applications with a likely beneficial outcome, and thus is likely not true. As a corollary, this Manifesto rejects the premise that computing is worth developing at any cost because it will help solve the world's problems. We assert that, with respect to solving increasingly pressing environmental problems, on the whole, computing as it is currently being developed is exacerbating rather than mitigating environmental and social harms.

3.2 Direct versus indirect impacts

Computing is not immaterial; it has numerous, multifaceted, and intertwined impacts on the environment. For this Manifesto, we adopt an established taxonomy that distinguishes between direct effects, beneficial indirect effects, and detrimental indirect ones [15].

Direct impacts occur through mineral extraction, the manufacturing, transportation, use and disposal of hardware that provides computation, storage, network access, and interactivity. These impacts include air/water pollution, GHG-emissions, water consumption, waste production, and mineral depletion – all of which are growing as computational-intensity increases and as new applications of computing are developed. These impacts arise through, e.g., fossil-fuel based energy powering of data centres, evaporative cooling, rare mineral extraction to create hardware, and largely illegal dumping and processing of electronic waste [57, 23, 31].

Self-reported GHG emissions from major tech companies are rising significantly due to the increasing development of deployment of generative AI such as large language models (LLMs) [8, 36, 47]. However, the true amount of energy used or emissions produced in the manufacture, use, and disposal of computing are unknown [18, 48]. There is an abundance of contradictory information regarding these direct impacts and, relatedly, a lack of transparency and disclosure from technology companies. It is not possible, for example, to understand the drivers of total energy consumption from the data that is voluntarily published; and recent investigations have claimed that the true emissions from data centres could even be 6–8 times higher than reports produced by the world's major tech companies [51]. Existing transparency requirements mandated by regulations such as in the EU AI Act [54] are focused on high-risk systems and impact incurred during model training. Inference reporting requirements [29] and supply chain disclosure are needed to develop more complete estimations of direct impacts and to understand opportunities for reductions, e.g. energy/emissions hotspots which can be more directly targeted. Transparent and standardised benchmarking is critical, and further work is needed to develop standards for measuring AI safety which includes energy requirements [65].

Indirect impacts arise from the transformative nature of computing, with effects across societal and economic sectors yielding both beneficial and detrimental impacts. Several mechanisms have been identified in Science and Technology Studies literature, notably how automation increases the resource intensity of everyday life through accumulation, acceleration, and stacking [41]. Typical bottom-up assessment methodologies for estimating indirect impacts are limited for a several reasons:

1. The ontologically uncertain set of mechanisms yielding indirect effects, and the epistemically uncertain assessment of those that are known.
2. The "chronic potentialitis" [21] of such assessments, which typically lie *in the future* and their occurrence is almost *never validated in hindsight*.
3. The plethora of different types of rebound effects [6, 60, 43, 22] that exist and can outweigh the positive indirect effects.
4. The difficulty in estimating the hypothetical baseline/counterfactual, often leading to its overstatement, which consequently also yields an overstated positive effect.
5. Possible time boundaries for indirect effects: when they become part of the socio-technical regime [35], should these effects no longer be considered additional?
6. The possibly difficult boundary between rebound effects and economic growth: are rebound effects merely one mechanism of economic growth, and if so, should they be counted as indirect effects of computing at all?

Top-down assessments, such as quantitative systems dynamics or input-output analyses, would address some of the first four limitations. As opposed to bottom-up assessments, they can set the system boundary arbitrarily wide and thus account for the subtle and hard-to-grasp mechanisms which end up being significant. For top-down assessments, however, causal

links are hard to establish, so they miss some of the explanatory power of bottom-up analyses. A hybrid approach deploying both might thus be called for, combining the useful aspects and compensating for the limitations of each. Crucially, we emphasise the importance of establishing a systematic framework of methods and standards for quantitatively assessing the impacts of various computing technologies, especially computationally intensive and fast growing ones, such as those due to generative AI, blockchain, and cryptocurrency.

3.3 Reducing impacts

Halting climate change requires that we commit to reducing the environmental impacts of computing systems: reducing direct, indirect, and other systemic impacts. Specifically, we must achieve *absolute reductions* in global GHG emissions caused by computing. To be compliant with the Paris Agreement, GHG reductions of around 7% per annum in the organisation and participation of Computer Science scientific and industry-oriented events are also necessary. And while other sectors bear responsibility for reducing GHG emissions, as computing researchers, professionals, educators, and users, we are not exempt and must commit to enabling the absolute reductions in global GHG emissions in any way possible.

To gain a clear understanding of whether and to what extent the above commitments are being met, we must also commit to sharing information regarding potential and actual impacts of computing in a transparent, evidence based, truthful, and holistic manner. Any claims regarding the beneficial impacts of computing must be substantiated *with evidence*. To this end, more comprehensive reporting is needed to assess the individual and comparative impacts of different computing solutions which are promising to reduce emissions, as well as the technological readiness of such solutions. There are already good overviews of what approaches can be taken (particularly for direct impacts) [42, 44, 45, 38, 48, 26], but there is a critical research-to-deployment gap and a comparative underdevelopment of approaches for assessing indirect impacts.

4 Shifting the Paradigm

In keeping with the principles of sustainability design as outlined in the Karlskrona Manifesto [12], a serious commitment to reducing digital technologies' environmental and social harms requires fundamental changes to the design priorities underlying hardware and software development. In contrast to environmentally destructive planned obsolescence, an anti-obsolescence paradigm would entail ensuring durability, longevity, repairability (e.g. modularity), and reusability of hardware. Incorporation of biodegradable parts should also be explored [17] (though should not be seen as a replacement for recycling parts wherever possible). Software should be low-resource and low-energy by design [32]. Anticipating the systemic impacts of climate change, it is also important to avoid catastrophic failures from critical infrastructures by developing technologies that are robust enough to withstand environmental or political instability (cf. "collapse informatics" [63]), i.e. ensuring adaptability, longevity, and resilience of software [32].[4] At the same time, a commitment to reducing demand necessitates a transition to self-obviating technologies and to supporting the development of skills that can, over time, decrease our technological dependence.

[4] Often these design aims are realised through lower-complexity technologies, and/or specialisation to specific tasks/use cases (rather than general purpose).

4.1 Prioritising sustainability

The current, unsustainable trajectory is a consequence of prioritising computational scaling over improved efficiency, or environmental sustainability[5] [23]. The only remedy is elevating sustainability to a first-class consideration for computing and computing facility design and implementation decisions. This means considering carefully where and how computing is deployed in the world – specifically, advocating for sufficiency in use of computational resources [61] and prioritising purposes that support sustainable development goals, human rights and scientific truths over uses that may support excessive consumption or antisocial behaviour. Crucially, this also means working to reduce demand for computing through cultivation of a reflective mindset regarding the purpose of the applications and services that drive demand (see, e.g., [56]). In some cases, the implication will be to apply less computationally intensive solutions, or to not employ computing at all for a given problem [10].

While decarbonisation (i.e. through use of renewable energy) is essential for reducing emissions arising from human activities which cannot yet be feasibly eliminated, it does not give licence to continue to create new and higher demand for computing technology. Decarbonisation of computing without demand reduction is not in line with a serious commitment to achieving absolute reductions in GHG emissions caused by computing; nor is it in line with a commitment to contributing to achieving absolute reductions in global GHG emissions. Renewable energy is limited by mineral availability / extraction capability and manufacture (all of which have associated environmental and humanitarian costs, including incurred GHG emissions), so using renewable energy to meet ever-rising computing demands diminishes the renewable energy available for decarbonisation of other sectors.

4.2 Within Planetary boundaries

Computing technologies do cause significant and varied social impacts, as well as environmental ones. Delivering positive social impacts necessarily entails environmental costs of the manufacturing, use, and disposal of the associated computing, and there is an implicit need to strike a balance between realising social good and maintaining a habitable planet. Traditional measures of human well-being, such as Gross Domestic Product (GDP), do not sufficiently account for this delicate balancing, ignoring the physical/material basis for sustained well-being.

We borrow the principles of Doughnut Economics [58]. The "doughnut" represents the "safe and just space for humanity" (ibid), between undershoot of social needs and overshoot of ecological limits. Applied to computing, this means contributing to technological innovation that helps humans meet longstanding sustainable development goals while upholding planetary boundaries [59], including but not limited to the limit of 2°C of global warming.[6] And for those boundaries that have already been surpassed, it means working to move back within a safe region, but doing so cognizant of the harms of deprivation (see "just sustainability" [11]).

---

[5] The term "sustainability" is deeply political, deployed variously depending on philosophical orientation and ideology [24]. This Manifesto recognises the climate change challenge as systemic in nature (section 5), involving the work of resolving contradictions between design goals and the political regimes into which they are deployed [28].

[6] Note that climate action is one of the sustainable development goals: "Goal 13: Take urgent action to combat climate change and its impacts" [50]. Our point is to underscore that sometimes these goals are in conflict.

4.3 Justice and human rights

We recognise that, as with other resources, computing is not equitably distributed today. While some use excessive amounts of computing resources, other do not have enough. We assert that working to support the equitable and democratic redistribution of computing capabilities is imperative. And, as a corollary, while environmental impacts of computing systems must be reduced overall, we acknowledge that when exploring possible trade-offs between direct and indirect impacts, in some cases a larger footprint might be needed to ensure decent living standards for all.

We recommend exploring, and where possible measuring, environmental and societal impacts, both, with social impacts included in reporting requirements. This would necessarily involve accounting for exploitative practices in supply chains, model training, and data labeling, as well as whether/how computing technologies help to advance human rights, well-being, and equality. We note that formal methods for assessing social impacts are conspicuously lacking. Too often, social benefits of computing technology are taken for granted, while detrimental social impacts are understood only after harm has materialised. This also necessitates a commitment to contributing to appropriate methods and measurement practices, both for environmental impacts and social impacts. It is crucial that pursuit of reductions of environmental harms of computing technologies is done in ways that advance human rights as defined by the Universal Declaration of Human Rights [7].

4.4 Cultivating a sustainability mindset

In circumstances of constraint (contrasting the current condition of unchecked growth in demand for computing), the efficiencies that computing naturally delivers could help drive GHG emissions reductions across the economy, thus enabling society to maintain a certain quality of life using less and less energy resources. But computing technologies can and do catalyse deeper changes to society. Hence, we echo others who have demanded that "Digitalisation must increasingly be put into the service of society and of a sustainable socio-ecological transformation" [9]. To this we add a commitment to help enable change by locating and acting on leverage points that are beyond the conventional scope of computing. And, to enable better, more sustainable, and more just decision making, it is essential that forecasting of environmental and social impacts are incorporated into all technological developments at an early stage. This should involve a thorough exploration of scenarios of use and potential indirect effects, e.g. considering what the consequences would be if a billion people used a given technology, or if a technology became critically entangled with another technology.

5 Additional Leverage Points

We recognise that limiting climate change is a systemic challenge. The climate crisis reflects the fundamental incompatibility of business-as-usual with the long-term life-sustaining health of our planet. Attending to this crisis requires changing high-level leverage points (e.g., societal structures and paradigms), even though it is difficult. Equally, computing professionals are not solely responsible for advancing computing, and its uses, cognizant of environmental limits and impacts. Governments, corporations, and educational institutions also bear responsibility, requiring unprecedented collaboration to effect the changes needed.

5.1 Influencing technology policy

Some form of regulation is needed to force a change to business-as-usual that will put us on a sustainable path. Individuals crafting regulation require good data to be able to make good decisions; but they also need to understand wider systemic effects such as rebound and the futility of efficiency-led solutions which call for ever more computing. Computing experts have an important role to play in informing and educating decision makers. Good science and solid numbers are needed to defeat discourses of delay [53]; hence, we reiterate the importance of exploring, and where possible measuring, environmental and societal impacts, and contributing to appropriate measurement methods and practices.

We also note that, as essential as it is to commit to leveraging computing to solve environmental challenges for climate solutions when relevant, it is also critical to counter techno-solutionism [16, 3], i.e. by advocating for policy/funding/innovation frameworks which ensure effective uptake and usage of computing solutions while minimising rebound effects. We contend that advocating for national and international digital policy that focuses on climate and environmental sustainability should be understood as a basic professional responsibility of those working in the field of computing.

5.2 Corporate truth and accountability

Historically, corporate reporting of impacts is prone to accounting that glorifies efforts that have the appearance of yielding environmental benefits (e.g. carbon offsetting) while minimising detrimental indirect effects. While on the one hand, this is a consequence of the immaturity of assessment methodologies and lack of standardisation required for meaningful accountability (as discussed above), it also highlights the dangers of transparency without the corresponding commitment to responsibility, honesty, and integrity from global businesses (cf. [25]). Much work is needed to develop robust accountability frameworks that prevent companies from bypassing their ethical responsibilities through strategic public relations efforts, and ultimately, from prioritising profits over ethical considerations.

We also note that the media today presents many outdated, contradicting, and false claims, making priority-setting and thoughtful policymaking difficult [53, 13]; and further, that social media has accelerated the spread of mis/disinformation. While there is no easy remedy to this problem, the more rounded assessment of environmental and societal impacts we advocate in this Manifesto would help reveal the interconnectedness of these issues. We stress, as well, the need to explore solutions for combatting mis/disinformation, including disincentivising (or, minimally, ceasing incentivising) its propagation through monetisation of engagement, and ending the domination of social media discourse by a small number of highly influential technology companies.

5.3 Revolutionising computing education

Shifting towards climate-conscious computing begins with cultivating a sustainability mindset through reforms to computing education. Accreditation requirements and curriculum development should reflect the changing skill set needed for professionals in a society undergoing socio-ecological transformation. Students should be taught to think about the consequences of the technology they will build and whether the technology is justifiable within a sustainability

framework, developing broader skills in systems thinking [30] and in engaging with social theory [55]. The implications for computing education are profound, and will require a more significant and coordinated effort to design, implement, and evaluate curriculum [55].

Likewise, we note the importance of building these same skills in those who have been working as computing professionals under the old paradigm. Professional computing bodies, such as ACM and IEEE, will need to play a role in advocating for new professional responsibilities entailed by our changing circumstances.

6 Conclusion

This Manifesto has highlighted the dual role of computing in climate change – that computing can, and is currently, driving a range of environmental harms, but that it can be a key driver for the positive change needed if it is more purposely steered towards enabling sustainable socio-ecological transformation. Modern science depends heavily on computing, and we will need more computing in the coming years to understand how to adapt to a changing climate. Growth in computing for environmental benefit requires, however, corresponding degrowth in computing that does not provide meaningful environmental or social benefit (e.g. addressing our most pressing sustainable development goals).

Our recommendations cover changes that need to be made to the design priorities of computing technologies, but also speak to the more systemic changes needed and their implications for computing. Among these, we emphasise that a radical shift in mindset, with sustainability and human rights providing the moral foundation for developing the kinds of computing technologies most needed by society, is imperative. We also stress the importance of digital policy that accounts for both the direct material impacts of computing and the detrimental indirect impacts arising from computing-enabled efficiencies, and the role of computing professionals in informing policy making.

7 Participants

- Christoph Becker
University of Toronto, CA
- Mike Berners-Lee
Lancaster University, GB
- Vinton G. Cerf
Google – Reston, US
- Andrew A. Chien
University of Chicago, US
- Benoit Combemale
University of Rennes, FR
- Vlad Coroamă
Roegen Centre for Sustainability – Zürich, CH
- Koen De Bosschere
Ghent University, BE
- Yi Ding
Purdue University – West Lafayette, US
- Adrian Friday
Lancaster University, GB
- Boris Gamazaychikov
Salesforce – Paris, FR
- Vicki Hanson
ACM – New York, US
- Lynda Hardman
CWI – Amsterdam, NL & Utrecht University, NL
- Simon Hinterholzer
Borderstep Institute – Berlin, DE
- Mattias Höjer
KTH Royal Institute of Technology – Stockholm, SE
- Lynn Kaack
Hertie School of Governance – Berlin, DE
- Bran Knowles
Lancaster University, GB
- Lenneke Kuijer
TU Eindhoven, NL
- Anne-Laure Ligozat
CNRS – Orsay, FR
- Jan Tobias Muehlberg
Free University of Brussels, BE
- Yunmook Nah
Dankook University – Yongin-si, KR
- Thomas Olsson
University of Tampere, FI
- Anne-Cécile Orgerie
CNRS – IRISA – Rennes, FR
- Daniel Pargman
KTH Royal Institute of Technology – Stockholm, SE
- Birgit Penzenstadler
Chalmers University of Technology – Göteborg, SE
- Chris Preist
University of Bristol, GB
- Tom Romanoff
ACM – New York, US
- Emma Strubell
Carnegie Mellon University – Pittsburgh, US
- Colin Venters
University of Limerick, IE
- Junhua Zhao
The Chinese University of Hong Kong – Shenzhen, CN

References

1 Earth's boundaries? *Nature*, 461:447–448, 2009. doi:10.1038/461447b.

2 Paris agreement, 2015. United Nations Treaty Collection, Chapter XXVII 7. d. URL: https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en.

3 Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 252–260, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3351095.3372871.

4 ACM. Acm europe council, 2025. URL: https://europe.acm.org/.

5 Blake Alcott, Mario Giampietro, Kozo Mayumi, and John Polimeni. *The Jevons paradox and the myth of resource efficiency improvements*. Routledge, 2012.

6 Elise Andrew and Daniela CA Pigosso. Multidisciplinary perspectives on rebound effects in sustainability: A systematic review. *Journal of Cleaner Production*, page 143366, 2024.

7 United Nations General Assembly. Universal declaration of human rights, 1948. Adopted 10 December 1948, Resolution 217 A (III), https://www.refworld.org/docid/3ae6b3712c.html.

8 Baidu. Baidu 2024 Environmental, Social and Governance Report. Technical report, Baidu, 2024. URL: https://esgs.cdn.bcebos.com/2025/04/28/a2/39a2aaa4f728f4e2d0aebaff346c045e.pdf.

9 Bits & Bäume. Making digitalisation sustainable and fit for the future: Political demands, 2022. URL: https://bits-und-baeume.org/en/conference-2022/demands/.

10 Eric P.S. Baumer and M. Six Silberman. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2271–2274, New York, NY, USA, 2011. Association for Computing Machinery. doi:10.1145/1978942.1979275.

11 Christoph Becker. *Insolvent: How to Reorient Computing for Just Sustainability*. MIT Press, 2023.

12 Christoph Becker, Ruzanna Chitchyan, Leticia Duboc, Steve Easterbrook, Martin Mahaux, Birgit Penzenstadler, Guillermo Rodriguez-Navas, Camille Salinesi, Norbert Seyff, Colin Venters, et al. The karlskrona manifesto for sustainability design. *arXiv preprint arXiv:1410.6968*, 2014. doi:10.48550/arXiv.1410.6968.

13 Mike Berners-Lee. *A Climate of Truth: Why We Need It and How To Get It*. Cambridge University Press, 2025.

14 Eli Blevis. Sustainable interaction design: Invention & disposal, renewal & reuse. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 503–512, New York, NY, USA, 2007. Association for Computing Machinery. doi:10.1145/1240624.1240705.

15 Christina Bremer, George Kamiya, Pernilla Bergmark, Vlad C Coroama, Eric Masanet, and Reid Lifset. Assessing energy and climate effects of digitalization: Methodological challenges and key recommendations. *nDEE Framing Paper Series*, 2023.

16 Christina Bremer, Bran Knowles, and Adrian Friday. Have we taken on too much?: A critical review of the sustainable hci landscape. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3491102.3517609.

17 Ting-Jung Chang, Zhuozhi Yao, Paul J. Jackson, Barry P. Rand, and David Wentzlaff. Architectural tradeoffs for biodegradable computing. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-50 '17, pages 706–717, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3123939.3123980.

18 Sophia Chen. How much energy will ai really consume? the good, the bad and the unknown. *Nature*, 2025. URL: https://www.nature.com/articles/d41586-025-00616-z.

19 Robert Comber and Elina Eriksson. Computing as ecocide. In *LIMITS 2023, Ninth Workshop on Computing within Limits, June 14-15 2023, Online*. PubPub, 2023.

20 Copernicus. Copernicus: 2024 is the first year to exceed 1.5°c above pre-industrial level, 2025. URL: https://climate.copernicus.eu/copernicus-2024-first-year-exceed-15degc-above-pre-industrial-level.

21 Vlad C. Coroamă. The chronic potentialitis of digital enablement, 2024. URL: https://www.linkedin.com/pulse/chronic-potentialitis-digital-enablement-vlad-constantin-coroam%25C4%2583-crtjc/.

22 Vlad C. Coroamă and Daniel Pargman. Skill rebound: On an unintended effect of digitalization. In *Proceedings of the 7th International Conference on ICT for Sustainability*, ICT4S2020, pages 213–219, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3401335.3401362.

23 Kate Crawford. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021. doi:10.2307/j.ctv1ghv45t.

24 Aidan Davison. *Technology and the Contested Meanings of Sustainability*. State University of New York Press, 2001.

25 Virginia Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, volume 2156. Springer, 2019. doi:10.1007/978-3-030-30371-6.

26 Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. Measuring the carbon intensity of ai in cloud instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1877–1894, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3531146.3533234.

27 Rodrigo dos Santos, Michelle Kaczmarek, Saguna Shankar, and Lisa P Nathan. Who are we listening to? the inclusion of other-than-human participants in design. In *Seventh Workshop on Computing within Limits 2021*. LIMITS, 2021.

28 Paul Dourish. Hci and environmental sustainability: The politics of design and the design of politics. In *Proceedings of the 8th ACM conference on designing interactive systems*, pages 1–10, 2010. doi:10.1145/1858171.1858173.

29 Kai Ebert, Nicolas Alder, Ralf Herbrich, and Philipp Hacker. Ai, climate, and regulation: From data centers to the ai act. *arXiv preprint arXiv:2410.06681*, 2024. doi:10.48550/arXiv.2410.06681.

30 Elina Eriksson, Miriam Börjesson Rivera, Björn Hedin, Daniel Pargman, and Hanna Hasselqvist. Systems thinking exercises in computing education: Broadening the scope of ict and sustainability. In *Proceedings of the 7th International Conference on ICT for Sustainability*, ICT4S2020, page 170–176, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3401335.3401670.

31 Vanessa Forti, Cornelis Peter Baldé, Ruediger Kuehr, and Garam Bel. The global e-waste monitor 2020. *United Nations University (UNU), International Telecommunication Union (ITU) & International Solid Waste Association (ISWA), Bonn/Geneva/Rotterdam*, 120, 2020.

32 Green Software Foundation. 2023 state of green software, 2024. URL: https://stateof.greensoftware.foundation/en/.

33 Charlotte Freitag, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S Blair, and Adrian Friday. The real climate and transformative impact of ict: A critique of estimates, trends, and regulations. *Patterns*, 2(9), 2021. doi:10.1016/j.patter.2021.100340.

34 Stephen M Gardiner. *A perfect moral storm: The ethical tragedy of climate change*. Oxford University Press, 2011.

35 Frank W Geels. The multi-level perspective on sustainability transitions: Responses to seven criticisms. *Environmental Innovation and Societal Transitions*, 1(1):24–40, 2011.

36 Google. Google 2024 Environmental Report. Technical report, Google, 2024. URL: https://sustainability.google/reports/google-2024-environmental-report/.

37 IEA. Global energy review 2025, 2025. URL: https://www.iea.org/reports/global-energy-review-2025.

38 Yunho Jin, Gu-Yeon Wei, and David Brooks. The energy cost of reasoning: Analyzing energy usage in llms with test-time compute. *arXiv preprint arXiv:2505.14733*, 2025. doi:10.48550/arXiv.2505.14733.

39 Bran Knowles, Kelly Widdicks, Gordon Blair, Mike Berners-Lee, and Adrian Friday. Our house is on fire: The climate emergency and computing's responsibility. *Communications of the ACM*, 65(6):38–40, 2022. doi:10.1145/3503916.

40 Lenneke Kuijer, Kirsikka Kaipainen, Nicola J Bidwell, Markus Fiedler, Adrian Friday, Marc Hassenzahl, Carine Lallemand, Matthias Laschke, Dan Lockton, Thomas Olsson, et al. Once upon a time when hci prioritised environmental sustainability: Reflections on a collection of fictions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 641:1–641:10, 2025. doi:10.1145/3706599.3716223.

41 Lenneke Kuijer and Matthias Laschke. Designing for a post-growth society through the eco-harmonist. a critical examination of the role of hci and technology design. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*, pages 1–13, 2024. doi:10.1145/3679318.3685405.

42 Benjamin C Lee, David Brooks, Arthur van Benthem, Udit Gupta, Gage Hills, Vincent Liu, Benjamin Pierce, Christopher Stewart, Emma Strubell, Gu-Yeon Wei, et al. Carbon connect: An ecosystem for sustainable computing. *arXiv preprint arXiv:2405.13858*, 2024. doi:10.48550/arXiv.2405.13858.

43 Alexandra Sasha Luccioni, Emma Strubell, and Kate Crawford. From efficiency gains to rebound effects: The problem of jevons' paradox in ai's polarized environmental debate. *arXiv preprint arXiv:2501.16548*, 2025. doi:10.48550/arXiv.2501.16548.

44 Sasha Luccioni. Light bulbs have energy ratings — so why can't ai chatbots? *Nature*, 2024. URL: https://www.nature.com/articles/d41586-024-02680-3.

45 Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 85–99, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3630106.3658542.

46 Jennifer C. Mankoff, Eli Blevis, Alan Borning, Batya Friedman, Susan R. Fussell, Jay Hasbrouck, Allison Woodruff, and Phoebe Sengers. Environmental sustainability and interaction. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, pages 2121–2124, New York, NY, USA, 2007. Association for Computing Machinery. doi:10.1145/1240866.1240963.

47 Microsoft. Microsoft 2024 Environmental Sustainability Report. Technical report, Microsoft, 2024. URL: https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report.

48 Jacob Morrison, Clara Na, Jared Fernandez, Tim Dettmers, Emma Strubell, and Jesse Dodge. Holistically evaluating the environmental impact of creating language models. *arXiv preprint arXiv:2503.05804*, 2025. doi:10.48550/arXiv.2503.05804.

49 San Murugesan. Harnessing green it: Principles and practices. *IT professional*, 10(1):24–33, 2008. doi:10.1109/MITP.2008.10.

50 United Nations. Sustainable development goals. transforming our world: the 2030 agenda for sustainable development. Technical report, New York, 2015. URL: https://sustainabledevelopment.un.org/topics.

51 Isabel O'Brien. Data center emissions probably 662it keep up the ruse? *The Guardian*, 2024. URL: https://www.theguardian.com/technology/2024/sep/15/data-center-gas-emissions-tech.

52 Intergovernmental Panel on Climate Change (IPCC). Climate change 2023: Synthesis report, 2023. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. URL: https://www.ipcc.ch/report/ar6/syr/, doi:10.59327/IPCC/AR6-9789291691647.

53 Naomi Oreskes and Erik M Conway. Defeating the merchants of doubt. *Nature*, 465(7299):686–687, 2010.

54 European Parliament. Eu ai act: first regulation on artificial intelligence, 2023. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

55 Anne-Kathrin Peters, Rafael Capilla, Vlad Constantin Coroamă, Rogardt Heldal, Patricia Lago, Ola Leifler, Ana Moreira, João Paulo Fernandes, Birgit Penzenstadler, Jari Porras, and Colin C. Venters. Sustainability in computing education: A systematic literature review. *ACM Trans. Comput. Educ.*, 24(1), feb 2024. doi:10.1145/3639060.

56 Chris Preist, Daniel Schien, and Eli Blevis. Understanding and mitigating the effects of device and cloud service design decisions on the environmental footprint of digital infrastructure. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1324–1337, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2858036.2858378.

57 Ramachandran Rajesh, Dharmaraj Kanakadhurga, and Natarajan Prabaharan. Electronic waste: A critical assessment on the unimaginable growing pollutant, legislations and environmental impacts. *Environmental Challenges*, 7:100507, 2022.

58 Kate Raworth. *Doughnut economics: Seven ways to think like a 21st century economist*. Chelsea Green Publishing, 2018.

59 Katherine Richardson, Will Steffen, Wolfgang Lucht, Jørgen Bendtsen, Sarah E Cornell, Jonathan F Donges, Markus Drüke, Ingo Fetzer, Govindasamy Bala, Werner Von Bloh, et al. Earth beyond six of nine planetary boundaries. *Science advances*, 9(37):eadh2458, 2023.

60 Miriam Börjesson Rivera, Cecilia Håkansson, Åsa Svenfelt, and Göran Finnveden. Including second order effects in environmental assessments of ict. *Environmental Modelling & Software*, 56:105–115, 2014. doi:10.1016/j.envsoft.2014.02.005.

61 Tilman Santarius, Jan CT Bieser, Vivian Frick, Mattias Höjer, Maike Gossen, Lorenz M Hilty, Eva Kern, Johanna Pohl, Friederike Rohde, and Steffen Lange. Digital sufficiency: conceptual considerations for icts on a finite planet. *Annals of Telecommunications*, 78(5):277–295, 2023. doi:10.1007/s12243-022-00914-x.

62 Douglas Schuler. Welcome to the cybercene? *Ninth Computing within Limits, June*, 14, 2023.

63 Bill Tomlinson, M. Six Silberman, Donald Patterson, Yue Pan, and Eli Blevis. Collapse informatics: Augmenting the sustainability & ict4d discourse in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 655–664, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2207676.2207770.

64 Kentaro Toyama. Technology as amplifier in international development. In *Proceedings of the 2011 IConference*, iConference '11, pages 75–82, New York, NY, USA, 2011. Association for Computing Machinery. doi:10.1145/1940761.1940772.

65 Carnegie Mellon University. Ai measurement science & engineering (aimsec), 2025. URL: https://www.cmu.edu/aimsec/index.html.

Manifesto from Dagstuhl Perspectives Workshop 24352

Conversational Agents: A Framework for Evaluation (CAFE)

Christine Bauer[*1], Li Chen[*2], Nicola Ferro[*3], Norbert Fuhr[*4], Avishek Anand[5], Timo Breuer[6], Guglielmo Faggioli[†7], Ophir Frieder[8], Hideo Joho[9], Jussi Karlgren[10], Johannes Kiesel[11], Bart P. Knijnenburg[12], Aldo Lipani[13], Lien Michiels[14], Andrea Papenmeier[15], Maria Soledad Pera[16], Mark Sanderson[17], Scott Sanner[18], Benno Stein[19], Johanne R. Trippas[20], Karin Verspoor[21], and Martijn C. Willemsen[22]

1 **University of Salzburg, AT.** christine.bauer@plus.ac.at
2 **Hong Kong Baptist University, HK, China.** lichen@comp.hkbu.edu.hk
3 **University of Padua, IT.** nicola.ferro@unipd.it
4 **Universität Duisburg-Essen, DE.** norbert.fuhr@uni-due.de
5 **TU Delft, NL.** avishek.anand@tudelft.nl
6 **TH Köln, DE.** timo.breuer@th-koeln.de
7 **University of Padua, IT.** guglielmo.faggioli@unipd.it
8 **Georgetown University, USA.** ophir@ir.cs.georgetown.edu
9 **University of Tsukuba, JP.** hideo@slis.tsukuba.ac.jp
10 **Silo AI, FI.** jussi.karlgren@silo.ai
11 **Bauhaus-Universität Weimar, DE.** johannes.kiesel@uni-weimar.de
12 **Clemson University, USA.** bartk@clemson.edu
13 **University College London, UK.** aldo.lipani@ucl.ac.uk
14 **imec-SMIT, Vrije Universiteit Brussel & University of Antwerp, BE.** lien.michiels@uantwerpen.be
15 **University of Twente, NL.** a.papenmeier@utwente.nl
16 **TU Delft, NL.** m.s.pera@tudelft.nl
17 **RMIT University, AU.** mark.sanderson@rmit.edu.au
18 **University of Toronto, CA.** ssanner@mie.utoronto.ca
19 **Bauhaus-Universität Weimar, DE.** benno.stein@uni-weimar.de
20 **RMIT University, AU.** j.trippas@rmit.edu.au
21 **RMIT University, AU.** karin.verspoor@rmit.edu.au
22 **TU Eindhoven & JADS, NL.** m.c.willemsen@tue.nl

Abstract

During the workshop, we deeply discussed what *CONversational Information ACcess (CONIAC)* is and its unique features, proposing a world model abstracting it, and defined the *Conversational Agents Framework for Evaluation (CAFE)* for the evaluation of CONIAC systems, consisting of six major components: 1) goals of the system's stakeholders, 2) user tasks to be studied in the evaluation, 3) aspects of the users carrying out the tasks, 4) evaluation criteria to be considered, 5) evaluation methodology to be applied, and 6) measures for the quantitative criteria chosen.

Perspectives Workshop August 25–30, 2024 – https://www.dagstuhl.de/24352
2012 ACM Subject Classification Information systems → Information retrieval; Information systems → Recommender systems; Computing methodologies → Natural language processing
Keywords and phrases Conversational Agents, Evaluation, Information Access
Digital Object Identifier 10.4230/DagMan.11.1.19

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Conversational Agents: A Framework for Evaluation (CAFE), *Dagstuhl Manifestos*, Vol. 11, Issue 1, pp. 19–67
Editors: Christine Bauer, Li Chen, Nicola Ferro, Norbert Fuhr

Dagstuhl Manifestos
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Acknowledgements We thank Schloß Dagstuhl for hosting us. Icons are taken from the Noun Project[1]. This work was partially supported by CAMEO[2], PRIN 2022 n. 2022ZLL7MW, and by the Excellence in Digital Sciences and Interdisciplinary Technologies (EXDIGIT) project, funded by Land Salzburg under grant number 20204-WISS/263/6-6022.

Executive Summary

This workshop brought together 22 experts from both academia and industry in neighboring areas – namely *Information Retrieval (IR)*, *Recommender Systems (RS)*, and *Natural Language Processing (NLP)* – in order to envision a new framework for the evaluation of *CONversational Information ACcess (CONIAC)* systems and to discuss the related research challenges.

The framework starts from the assumption that a CONIAC system will be able to (i) *interact* with users more naturally and seamlessly, (ii) guide a user through the process of *refining* and *clarifying* their needs, (iii) *aid decision-making* by making *personalized recommendations and information* while being able to *explain* them, and (iv) *generate*, *retrieve* and *summarize* relevant information. To this end, a CONIAC system needs to cross the boundaries and mix different technologies, e.g., IR and RS, rely on both internal knowledge about the user and external knowledge about the context where it is operating, and, finally, track the stream of conversation events and dynamic changes in the user's belief and information state.

To support the development of such CONIAC systems, the *Conversational Agents Framework for Evaluation (CAFE)* recognizes that, to follow the flow of conversational events and system states, evaluation needs to consist of a dynamic *sequence of evaluation probes*, rather than being a single static assessment as it is today. Then CAFE introduces six areas to guide researchers and developers in defining what the evaluation probes should be for the case at hand:

- **Stakeholder Goals** discusses the diverse and often implicit objectives of various stakeholders involved in system design. These stakeholders include users, system designers, app developers, distribution platforms, content creators, publishers, advertisers, and editors. Each group has distinct goals that may overlap or conflict with others. For a system to be successful, it must address and balance these differing objectives. The section also notes the importance of enriching user interactions by supporting secondary goals, such as relationship-building and trust assessment, which enhance the overall user experience.
- **User Aspects** investigates the different angles of users approaching systems with diverse needs, such as answering questions, gaining knowledge, planning trips, or asking for recommendations. Some users or user contexts might benefit more from conversational systems due to some characteristics such as impairments or modality (e.g., hands-free operation). Users' conversational information access tasks can also vary based on cultural contexts, affecting how they interact with systems. When evaluating the system, it will be essential to consider these user aspects, as different users may have very different conversations and experiences with the system.

[1] https://thenounproject.com/
[2] https://cameo.dei.unipd.it/

- **Task** explores the characteristics and tasks suitable for CONIAC systems, which are defined by factors such as information need, human involvement, goal orientation, and iterative interaction. The proposed model categorizes tasks based on their complexity and definiteness, with well-defined tasks needing focused iterative interactions and ill-defined tasks requiring exploratory search and learning. This section presents examples, such as hypothesis formulation, product search, travel planning, and health information seeking, illustrating how CONIAC systems can adapt to evolving user needs and provide more tailored assistance compared to conventional systems. The discussion extends to emerging LLM-based conversational enterprise and personal information management.
- **Criteria** discusses the criteria to consider when instantiating an evaluation framework for the conversational search scenario. Such criteria can be organized into a taxonomy according to who the subject of the evaluation is. The first layer of this taxonomy divides the criteria into system- and user-centric criteria. The former concerns the system and can be assessed in isolation in a purely offline setting. The role of the human, in this case, is to provide labels. The latter regards user experience and the consequences associated with the usage of the system. These top-level criteria are then further specialized along several dimensions.
- **Methodology** organizes evaluation methodologies – both quantitative and qualitative – along two dimensions: (1) according to the focus of a study (a standard dimension in IR), ranging from system-focused methodologies like offline simulations to user-focused methodologies like qualitative interviews; and (2) according to the employed time model (a dimension from system behavior theory that is especially suitable for CONIAC studies), ranging from stationary methodologies like single-interaction experiments to methodologies like longitudinal user studies that allow for continuous measurements of, for example, satisfaction. Indeed, while evaluation criteria are defined as agnostic of time models, for the actual evaluation one has to decide which time model to use, which limits the available evaluation methodologies and affects the details of their implementation.
- **Measures** presents the commonly used metrics for assessing system-centric hardware and software criteria, and the measures for evaluating user-centric criteria (including users' subjective reflection on their interactions, their perceptions, and behavioral metrics). Depending on the goals and criteria a specific system aims to fulfill, one can adopt the suitable evaluation methodology (see examples in Table 2) to assess the corresponding metrics/measures.

When designing an evaluation, the first step is to identify the stakeholders and their goals that need to be addressed. Based on the goals, the user tasks to be studied in the evaluation have to be defined, and the user aspects need to be considered. The central elements of an evaluation are the criteria to be focused on, which can be determined by the stakeholder goals. The chosen criteria restrict the range of possible evaluation methods (e.g. any user-centric criterion requires the involvement of actual users in the evaluation procedure). Finally, an appropriate measure has to be defined for any quantitative criterion.

Overall, the CONIAC world model and the CAFE framework propose a series of "*high risk, high gain*" research topics that promise to deliver a major paradigm shift in the field of conversational agents. By embracing a new vision for conversational information access and targeting a technological breakthrough, these initiatives aim to transform how academia and industry invent, design, and develop such systems.

Table of Contents

1 Introduction

Every day, individuals interact with a wide variety of *agents*: web search engines to find information, recommender systems to identify content or products to purchase, or chatbots for customer service inquiries. Given their prevalence, let us focus on *conversational agents*, those that mimic human conversation using either text or spoken language, and more particularly on those enabling *information access*. While the literature in this area is rich [133], the evaluation of these systems is still limited to simplistic approaches that mostly ignore the conversational nature of the human-system interaction. In this manifesto, we specially focus on information access systems [105, 126] which allow us to seek, access, and use relevant information.

The very idea of some conversational interaction with information access systems traces back several decades. For example, Codd [29] early imagined conversations as a key way to allow casual users to interact with databases in a natural way, avoiding the complexity of a formal query language. Conversations figured out by Codd were very close to the idea of conversation we are still talking about nowadays, e.g., Codd [29, p. 182]:

```
USER: Give me a list of the part numbers, quantities,
      and suppliers of Houston parts

SYSTEM: You are requesting (among other thing) a listing of
        quantities of parts. Do you want:
        1. quantities on hand
        2. quantities on order
        3. quantities shipped?

USER: I want quantities on order

SYSTEM: Fine! Now, with regard to "Houston parts" do you mean:
        1. parts supplied by Houston supplier
        2. parts supplied to Houston projects
        3. parts associated with Houston in some other way?

USER: I mean item 2
```

The above conversation reveals the assumption of querying a relational database and the need to specify which relations/tables to query, possibly joining them, and which attributes within those relations/tables.

Recent years have witnessed a spurring of research on conversational agents, in multiple intersecting domains, such as *Artificial Intelligence (AI)* [46], *Information Seeking (IS)* [133], *Information Retrieval (IR)* [67, 7, 70], *Recommender Systems (RS)* [61], *Natural Language Processing (NLP)* [97, 80], and even more with the rise of *Large Language Models (LLMs)* and *Retrieval-Augmented Generation (RAG)* [5].

A natural question arises after so many years of research by multiple disciplines: *"Isn't this a solved problem?"*.

In order to answer this question, in Fig. 1, we present an example of how we imagine future conversations between humans and agents, inspired by Di Noia et al. [38]. We choose product search, a well-explored domain where it should be easier to answer "yes, it is a solved problem!" Differently from the early conversation imagined by Codd [29] and confined to the

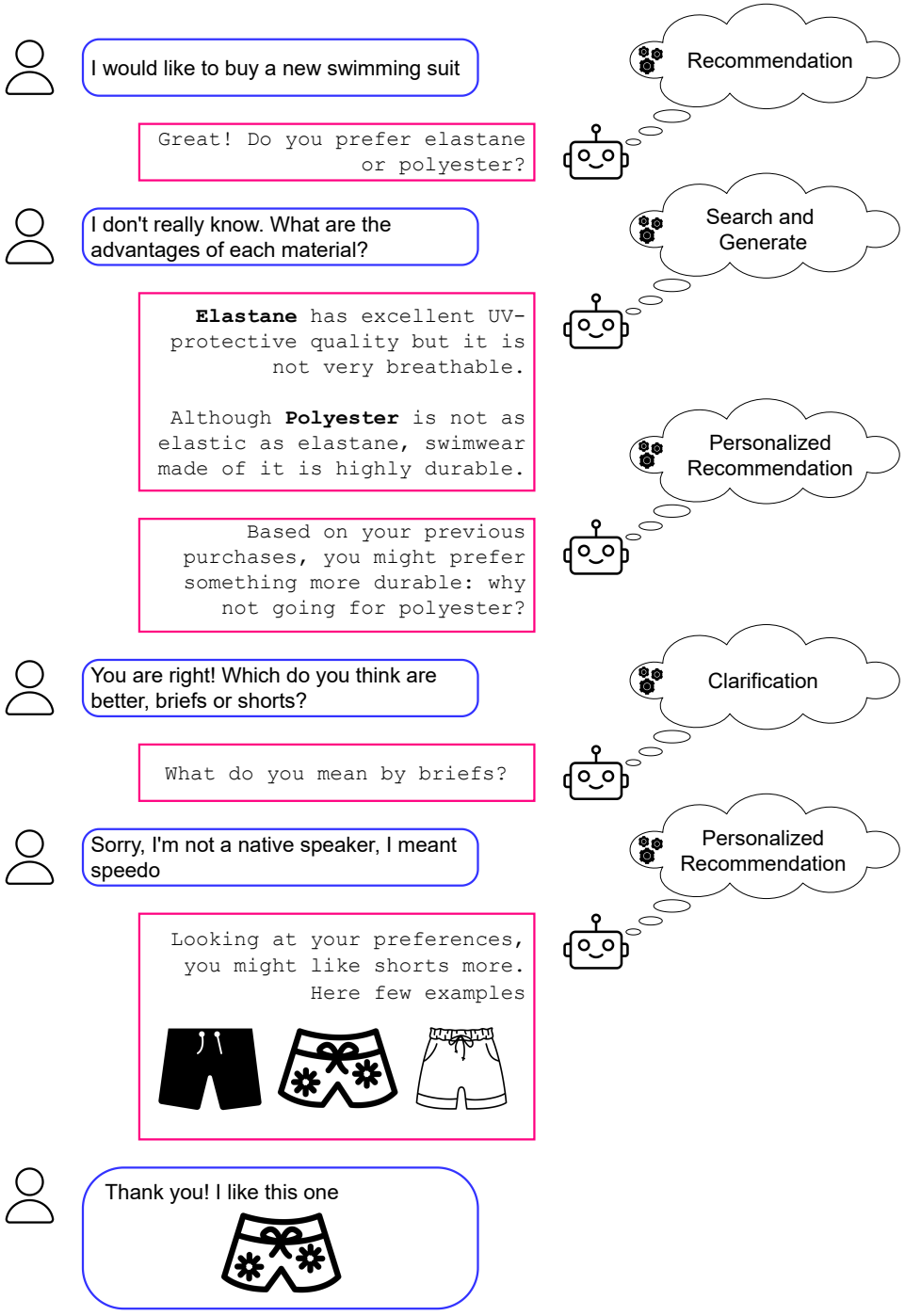

Figure 1 An example of a possible conversation in the product search domain.

realm of relational databases, in our case, the conversation naturally spans different types of systems and methodologies – e.g., IR systems and RSs – even if some commonalities are there, such as the system asking for clarifications when needed.

In the example, the user wants to buy a new swimsuit, but he is initially unsure or unaware of his exact requirements for this swimsuit. To support his decision-making process, the user initiates a dialog with the system: "I would like to buy a new swimming suit." The system

understands that the user's information need is best served by a RS component. Therefore, it responds with an attribute-related question: "`Great! Do you prefer elastane or polyester?`" Since the user does not know the pros and cons of the two materials, he asks for an explanation of the system: "I don't really know. What are the advantages of each material?" Initially, the system relies on a combination of IR and generation, e.g., a RAG component, to answer the user information need; afterward, it relies on a personalized RS to suggest going for polyester, since this material best matches with the user preference for more durable items. The user appreciates the suggestion and goes on asking "Which do you think are better, briefs or shorts?" Since the system does not correctly understand the term "briefs", it relies on a clarification component to ask for an explanation, "`What do you mean by briefs?`" Once the user clarifies that he actually means "speedo", the system relies on a personalized RS component to suggest some possible shorts – a type of swimsuit that better fits with the user preferences and profile – and, finally, the user picks one of them, ending the conversation. Fig. 1 also shows that, while we typically think of conversations made up just of textual utterances, we should generally consider *multimodal* and *multimedia* conversations, where more complex phenomena may happen as, e.g., interruptions in case of a spoken interaction; in this simple example, this fact is represented by images of swimsuits in some utterances.

To assess the extent to which this hypothetical conversation can be already served by current technology, we carried out a toy (and qualitative) experiment and ran this simplistic conversation, i.e., the sequence of utterances issued by the user of Fig. 1, using both GPT-3[3], shown in Fig. 2, and GPT-4 via Microsoft Co-pilot[4], shown in Fig. 3. The experiment was run in mid-October 2024 and, obviously, has several limitations, among which is the fact that GPT-3 and GPT-4 neither are backed by a RS nor do these have a specific user profile that can be used for personalization. Nevertheless, the type of answers returned gives an impression of how much of the problem is solved.

The example conversation with GPT-3 in Fig. 2 highlights some limitations. First, the generated answers tend to be quite lengthy, more than what a user might expect, and they are perhaps schematic, tending to rely a lot on item lists; on the positive side, utterances typically include a final recap. Obviously, GPT-3 cannot follow up the comparison among different materials with a personalized recommendation. GPT-3 does not perceive any ambiguity between `briefs` and `speedos`, while native speakers might do, and the clarification by the user produces a slightly more adapted response. On the positive side, GPT-3 maintains the context between the two utterances and continues to compare `shorts` and `speedos`. On the negative side, we can observe as in all the responses, the dimensions of comparisons are somewhat inconsistent since, for example, when it comes to materials they consist of entirely different facets or when comparing `briefs` vs. `shorts` and then `speedos` vs. `shorts` some explanations move from one category to another, e.g., freedom of movement is under mobility for `briefs` and under fit for `speedos`. Again, GPT-3 cannot complete the interaction with a personalized recommendation as in the case of Fig. 1.

The example conversation with GPT-4 in Fig. 3 highlights some limitations. On the positive side, GPT-4 strives to provide references to support its answers. On the negative side, the answer to the first utterance is very much focused on just recommending shops or brands rather than exploring the actual information need of the user. Moreover, the answer itself is a bit mixed and not fully homogeneous because some of the suggestions

[3] `https://chatgpt.com/`
[4] `https://copilot.cloud.microsoft/`

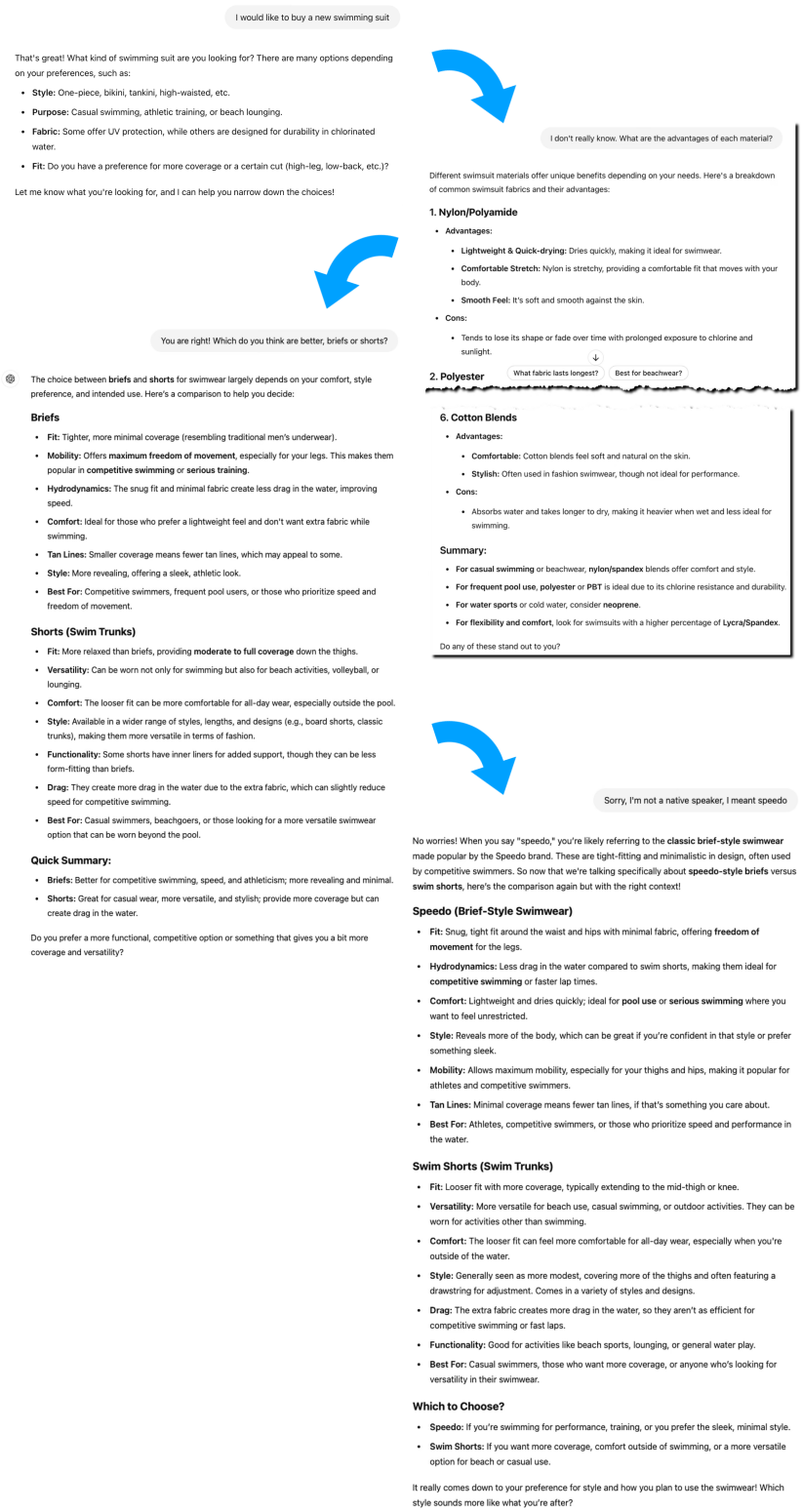

Figure 2 An example of conversation of Fig. 1 using GPT-3.

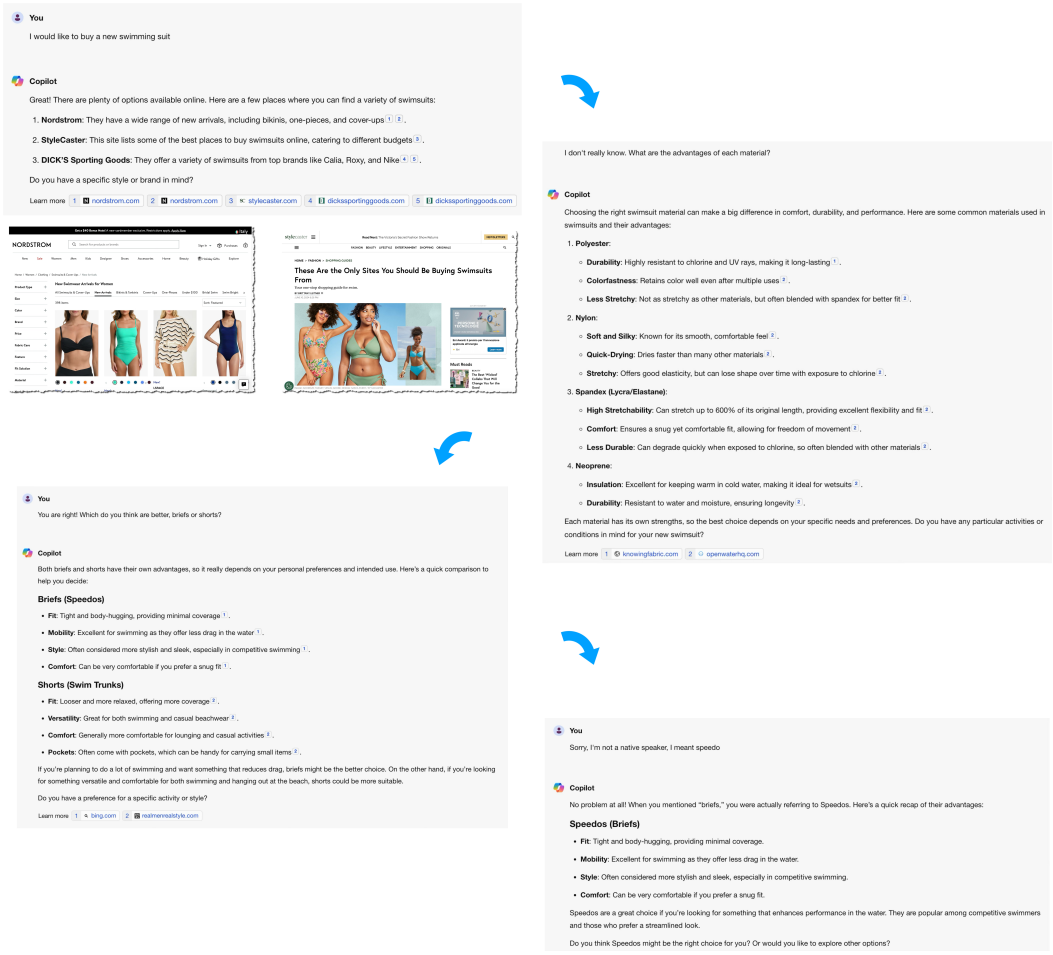

Figure 3 An example of conversation of Fig. 1 using GPT-4 via Microsoft Copilot.

point to e-commerce websites (e.g., Nordstrom) while others are fashion blog posts (e.g., StyleCaster); this somewhat recalls the issue mentioned above for GPT-3 about the dimensions of comparison being inconsistent. When it comes to comparing alternatives, the answers are much more compact and to the point than in the case of GPT-3, still suffering, however, from inconsistent dimensions across different materials. As with GPT-3, GPT-4 is not capable n of following up the explanation of materials with a personalized recommendation. When it comes to the comparison between `briefs/speedos` vs. `shorts`, GPT-4 explicitly considers them as synonyms, while a native speaker might not do so. If we force GPT-4 to answer just about `speedos`, on the positive side, the answer remains fully consistent with what replied in the case of `briefs/speedos` but, on the negative side, context is completely lost, and no comparison to `shorts` is made anymore. Also, GPT-4 does not complete the interaction with a personalized recommendation.

If we compare the interaction style of GPT-3 with that of GPT-4, apart from observing the more succinct answers in the latter case, we can also see that GPT-4 is very much focused on somehow narrowing down the scope of the conversation because the initial attempt is to immediately propose products, rather than exploring the actual needs of the stakeholder and asking for clarifications.

Overall, while the LLMs and chatbots of today are advanced systems, already performing well, they are currently insufficient to fully embody the vision (outlined above) of conversational interaction with information access systems, even in a well-studied and relatively confined domain such as product search and recommendation.

It is not hard to imagine more complex information access challenges, such as finding an object described by both structured data and free text elements. We could imagine someone asking a system to find a walk to do one evening. The system could consider walks within the user's driving, cycling, or public transport range and reply with suggestions of different levels of difficulty, length, and ascent. The searcher could then ask the system to find a subset of walks where reviews have described sunset vistas. Here, the system might draw such information from multiple reviews. The system could also incorporate elements of recommendation by looking at the searcher's walk history and employing some form of collaborative filtering to identify the ideal evening walk further.

The ideal system is likely to be even more complex. When one uses a traditional web search engine like Google, the system is actually a constellation of search services, each with its own searching approach: web search retrieves authoritative content, news retrieves from high-quality sources with a bias towards recency, image search will consider image quality and size, etc. Upon receiving a query, the search engine chooses which blend of services will best serve the query. A conversational information access system will likely need to provide the same suite of information access services that one sees in traditional search, as well as a way to identify which service best serves a particular conversation.

Therefore, we argue that we need to imagine a new generation of systems – we call them *CONversational Information ACcess (CONIAC)* systems and introduce them in Section 2 – able to adopt a more holistic approach to conversational interaction. Above all, and this is the primary focus of this report, we need to envision a new evaluation framework – we call it *Conversational Agents Framework for Evaluation (CAFE)*, we introduce it briefly in Section 3, and then expand on its details in sections 4 to 9 – capable of supporting and driving the design and development of CONIAC systems.

2 *CONversational Information ACcess (CONIAC)*

Fig. 4 depicts the *CONversational Information ACcess (CONIAC) World Model*, where we take a very broad view by considering any context and system that addresses a user's information need using language interactions. Indeed, given the wide range of domains and tasks that can prompt conversational events – each with its inherent peculiarities – it is impractical to exhaustively define every information discovery use case that might benefit from conversational aspects. Therefore, in the rest of this section, we introduce a layered abstraction – the CONIAC World Model – that allows us to represent the conversational process targeting information access from various perspectives.

Our broad view encompasses aspects of conversational information access systems as defined by Balog [11]:

> [W]e use the term conversational information access (CIA) to define a subset of conversational AI systems that specifically aim at a task-oriented sequence of exchanges to support multiple user goals, including search, recommendation, and exploratory information gathering; that require multi-step interactions over possibly multiple modalities. Further, these systems are expected to learn user preferences, personalize responses accordingly, and be capable of taking initiative.

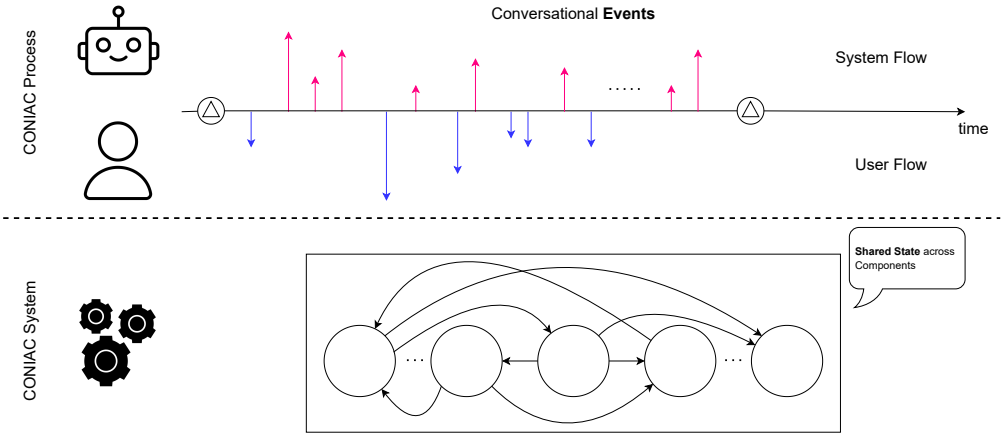

Figure 4 The CONIAC World Model.

and conversational information seeking as defined by Zamani et al. [133]:

> *Conversational Information Seeking (CIS)* is concerned with a sequence of interactions between one or more users and an information system. Interactions in CIS are primarily based on natural language dialogue, while they may include other types of interactions, such as click, touch, and body gestures.

While we take a broad view of CONIAC systems, one that encompasses both traditional conversational IR and RS, we envision that CONIAC will be most useful in interactive scenarios when users have complex or poorly understood information needs.

The CONIAC World Model consists of two layers: the CONIAC Process Layer (Section 2.1), where the actual conversation happens, and the CONIAC System Layer (Section 2.2), which is what we aim to design and develop.

2.1 The CONIAC Process Layer

We use the term *CONIAC Process Layer*, or **conversational process** for short, to refer to the overall conversation flow, i.e. any interaction between a CONIAC agent (the System Flow) and a human (the User Flow), via the interface of the system in a *multimedia* and *multimodal* way, in the most general setting. Some examples of conversational processes and mechanisms that influence the actual moves, utterances, or turns being generated are those that establish and maintain interpersonal rapport; those that organize the interaction channel through turn-taking cues, meta-communicative turns and even non-verbal actions such as gestures, facial expressions, and laughter; those that structure the informational flow of the conversation and narration as it proceeds, manage continuity and coherence, and negotiate topical shifts or switches when necessary; and those that define, modify, and address discourse referents and their properties. These include turns such as plain utterances, clarification questions, greetings, repairs, and verifications [53, 49, 54, 123]. Many conversational mechanisms are intuitively acquired by humans without instruction, though some are more explicitly trained in education or professional contexts.

More specifically, and more related to information access, conversational processes of interest include system and user collaborating in the expression of an information need, revealing the system's capabilities and collection qualities, clarification questions using mixed initiative, and persistence of discourse referents, allowing both parties to refer back to recently mentioned items or sets; and the ability to reason about the qualities and utility of a set and to further refine it for the purpose at hand [96].

An important assumption we make is that every conversational process addresses a single information need and thus has a single objective. This is a necessary assumption for our framework as, later on, this objective informs the selection of aspects to be evaluated, e.g., which criteria or measurements to use.

Specifically, a conversational process comprises different **conversational events** that occur over continuous time. We mean event in the most abstract sense, ranging from simple and atomic textual utterances to complex and *compound multimedia* and *multimodal* interactions, where phenomena such as interruptions, which for instance is central to spoken interaction, may happen. We distinguish four main conversational events – start, end, user, and system events – , that will be described in greater detail in the following paragraphs.

We further assume that an arbitrary amount of time may pass between two conversational events. Importantly, this includes possible interruptions, breaks, or pauses, after which the user or system decides to resume the conversation. This requires that the "state" of the conversation is maintained for arbitrary lengths of time. While concrete implementations may differ concerning the duration for which the state of the conversational process is preserved, we assume here that users may abandon and return to a session at any time without loss of generality.

Firstly, we make explicit that every conversational process starts with a **start event** and ends with a **end event**. Between this start and end event, the conversational process is comprised of "system" and "user" events, both of which correspond to language (or more general) interactions, either issued by the system (system events) or the user (user events). Importantly, we do not require that conversational processes alternate between user and system events. Users may issue several "user events" in a row, and so may the system. This is exemplified, for instance, in Fig. 1 where in response to the user event, "I don't really know. What are the advantages of each material?", the system issues two events, one explaining the advantages of each material and the other to perform a personalized recommendation.

Importantly, we distinguish start and end events from other types of conversational events. This is because the start of the conversational process can be a user event – e.g., "I need to learn about Hillary Clinton's career as a Secretary of State for a test. What do I need to know?" – or a system event – "Hello, my name is Sven. I am here to help you. What would you like me to help you with?" – , but also other types of non-language interactions between the user and the system, e.g., the user opening the chat window. The same is true of the end event.

Despite its apparent simplicity, this abstract and event-based view of the conversational process is flexible and powerful enough to be an instantiated model and represents all the conversational mechanisms and information discovery use cases briefly exemplified above.

2.2 The CONIAC System Layer

Unlike general interactive information systems, a CONIAC system must deal with the conversational dynamics described above, in particular by taking or relinquishing the initiative when necessary to keep the conversation flowing naturally [123]. In this regard, CONIAC

systems differ from many other systems because the latter often have one-shot interactions or repeated interactions that can be viewed in isolation. An example of the latter could be repeated query reformulation when using a search engine, e.g., "Hillary Clinton", followed by "Hillary Clinton state secretary", where each time the IR system responds with a list of documents relevant to the current query and there is no notion of "state" from one query to the next. In contrast, in our process-oriented view of CONIAC systems, this notion of *state* is essential, as it allows the system to use knowledge of past interactions to refine current and future interactions.

The bottom part of Fig. 4 depicts the *CONIAC System Layer*. We make abstraction of real-world CONIAC systems and only assume that the system will emit events that are generated by one or more components of the system, e.g., IR or RS components, essentially blurring the boundaries among components. These events can be, for example, responses to information needs, suggested items, explanations, or requests for clarification from the user. Note that by CONIAC system, we refer to both the underlying application and its interface.

As mentioned previously, another important assumption we make as a result of our view of a conversation as a process requires that there is some *shared state* across interactions between the user and the system, but also between different components – IR, RS, *Question Answering (QA)* – that together make up a CONIAC system. A state may even be recorded across different sessions – conversational processes – with the system, resulting in personalized interactions tailored to the user's (interaction) preferences.

The system layer is critical for managing and guiding interactions in the conversation process layer. We envision the system layer as implementing a stateful machine capable of flexible/expressive conversation flow tracking, personalization, and information access functionalities. The following functionalities are, in our view, essential for handling complex tasks involving information access:

- **World knowledge.** Reflects the external environment or domain-specific knowledge the system operates within, enabling it to provide contextually appropriate responses.
- **User information.** Incorporates user preferences or utilities, goals, and potentially private user-specific information, facilitating personalized interactions.
- **State tracking.** This state tracking involves the context that tracks the flow (both user and system events defined in Section 2.1) and the context of the conversation, including dynamically updated beliefs in the user information state revealed by user events.

State tracking in this scenario ensures that the CONIAC system can proactively generate system events – such as recommendations, clarification questions, or acknowledgments – in a manner that is informed by both the ongoing conversation and the user's inferred goals and preferences. Such system events should be generated in light of a range of objectives – e.g., coherence and relevance throughout the conversation, task completion, and efficiency – as specified by the system designer.

As an example of a possible concrete instantiation of a CONIAC system, Boutilier [18] defines a preference elicitation recommendation conversation as a *Partially Observable Markov Decision Process (POMDP)* model, which additionally requires a *reward model* to encapsulate the agent's interaction objectives. One advantage of such a decision-theoretic system model (which may be optimized through reinforcement learning methods when exact model details are unknown) is that it provides algorithmic and learning methodologies for optimizing system actions concerning the system design. This stands in contrast to many existing conversational dialog systems leveraging rule-based decision-making that may not optimize an explicitly defined system objective.

2.3 Limitations of Current Systems

Limitations of Information Retrieval Systems

Traditional IR systems are particularly limited in their capacity to handle these scenarios because although the queries are ambiguous, the assumption in most cases is that users' information needs are relatively well-defined from the outset. These systems are primarily designed to interpret and respond to explicit queries, focusing on accurately understanding and estimating a pre-existing, well-formed information need [26, 23, 8]. Consequently, they are less effective in situations when a user's needs are still evolving or are only partially articulated, which is common in many conversational contexts [68]. In such scenarios, the rigid query-response paradigm of conventional IR systems falls short, as it does not adequately support the dynamic, iterative process of refining and clarifying user intent that is essential for effective conversational information access.

Interactive Information Retrieval (IIR) systems are designed to support and even encourage user exploration within the information-seeking process, yet they are fundamentally limited by the fact that interaction is always initiated by the user. In these systems, the system consistently assumes the role of a passive responder, reacting to user queries rather than actively contributing to the dialogue. While such systems can effectively address well-defined user intents, they often fall short in scenarios where the user may have a clear goal but lacks awareness of all relevant aspects, possibilities, and available design or choice options [128, 85].

This limitation means that users might not fully explore the breadth of information or decision-making options available to them, as the system does not proactively guide them toward uncovering unknown preferences or considerations. As a result, the interaction remains user-driven, potentially leading to missed opportunities for discovery and a less comprehensive understanding of the subject matter. To overcome these challenges, more advanced systems that can take on a more active role in the interaction, such as by prompting users with relevant questions, suggesting alternative pathways, or highlighting overlooked options, are needed. These systems would better support users in navigating complex information landscapes and making more informed decisions [66, 15, 128, 85].

Limitations of Recommender Systems

Recommender Systems (RS), while highly effective in providing personalized content and suggestions based on user preferences, are also limited when applied to conversational scenarios. RS are inherently push-based systems, where suggestions are proactively presented to the user based on a pre-existing model of user behavior, preferences, and past interactions. Although they are beneficial in providing insights about choices, aspects, and actions that are absent in IR systems, recommender systems often lack query understanding capabilities. They do not interactively interpret or refine user intent based on evolving dialogue. Finally, both the IR and the RS approaches are mostly one-shot in the sense that they are optimized for single interaction settings. Even in session-based search [34] reformulations to an initial query result to a one-shot retrieval at the end of a session, and do not allow for intent modification or asynchronous change in intent.

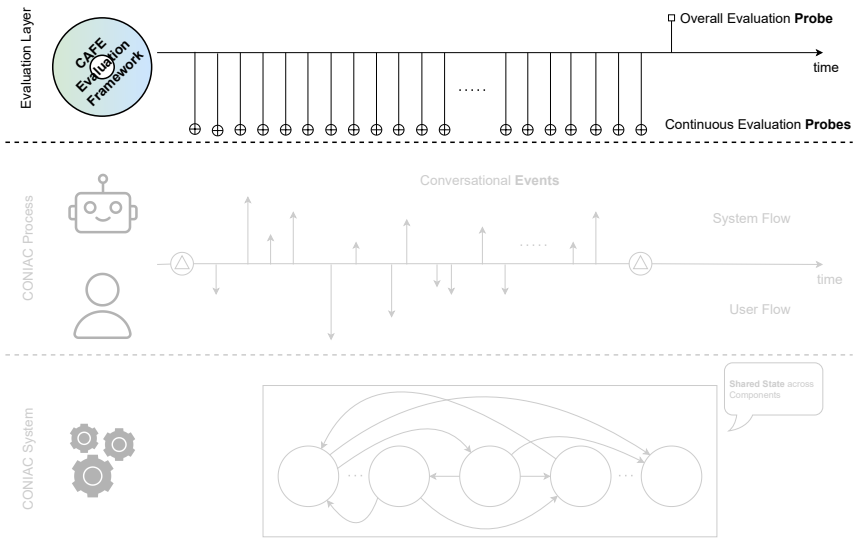

Figure 5 Evaluation Layer in the CONIAC World Model.

3 *Conversational Agents Framework for Evaluation (CAFE)*

In the world model depicted in Fig. 4, we outline a generic representation of the conversational scenario. To evaluate in this context, we need to understand how and when measurements are made to validate the goodness of the system that guides information discovery. Goodness can be a complex concept to capture in a single measure for evaluation, as many facets can directly influence it. Consequently, we argue that depending on the use cases, CONIAC systems are expected to encompass a wide range of different objectives and criteria for various stakeholders.

Therefore, in order to explain the overall Conversational Agents Framework for Evaluation, we first introduce the notion of an *Evaluation Layer* (Section 3.1) in the CONIAC world model and, then, we explain of CAFE (Section 3.2) allows for a concrete instantiation of such layer for actual evaluation purposes.

3.1 The Evaluation Layer

The vision of the CONIAC World Model depicted in the previous section calls for a paradigm shift in evaluation. As shown in Fig. 5, we can imagine that the *CONIAC Evaluation Layer* sits on top of the CONIAC process and system layers, to allow for their assessment, where the CONIAC system is the *construct* under evaluation.

The literature is rich with examples of how to approach the evaluation problem from multiple perspectives [78, 83, 59, 82, 93]. However, in general, these assessment strategies focus on a more static approach, focused mostly on overall performance with the conversational system or user satisfaction with (their interactions with) said system.

We argue that when it comes to CONIAC, evaluation should be inherently **dynamic** and **multi-faceted**. Indeed, we are aware that there is no optimal set of measurements and measures that can, for every CONIAC use case, comprehensively monitor the system

(Section 2.2), the users (Section 2.1), and the constantly evolving exchanges that take place. Consider an example of a direct answer to an unambiguous question "*what is the weather like*". This direct answer might be assessed in terms of fast-time response and correctness (e.g., the system can provide the right information for the right location). In contrast, for a more complex inquiry (e.g., a child explores to understand better a new concept ("*Can animals sleep upside down*"), assessment should be more focused not just on the ability of the system to understand the information needed and provide a relevant response, but also on other perspectives inherent to this particular user. In this example, perspectives could include the ability of the system to provide suitable information and to respond in a manner that a child can comprehend [78, 6]. However, a clear pattern emerges: regardless of "what" (i.e., criteria) is assessed or "how" (i.e., measures), assessment of CONIAC systems cannot be done in a summative manner, based on a single measurement.

With that in mind, we envision an evaluation layer that allows simultaneous and **continuous probing** of users, systems, and their interactions from multiple perspectives. Note that probes can happen synchronously with or asynchronously from the events in the conversational process, allowing for a wide range of evaluation styles.

In this layer, we define **probes** that act as sensors that constantly capture snapshots of the actions and reactions inherent to the conversational process from the start of the conversation all the way to the end (regardless of the resolution). Probes can be as simple as a single-valued measurement or complex multimedia objects, e.g., multi-valued measurements together with log records and video recordings of the user interaction. The idea is not to rely on a single probe, but instead – depending upon the use case – use a variety of probes that capture the different facets of the conversational process to assess.

The continuous probing is what then makes it possible not only to yield an overall measurement for the system – it can be some form of aggregation of the collected probes or an independent probe on its own – but also to gauge the correctness (or lack thereof) of the system events, along with user events. The measurements captured by these probes can also yield cumulative measurements, in addition to providing snapshots of the behavior of the different components of the system (and users' reactions to those).

Given that the probes can encode a rich and diverse set of measurements, it is possible to simultaneously capture measurements related to effectiveness and user behavior (via, for example, eye tracking), timeliness, and more.

Note that we focused our discussion on how evaluation happens as a form of external assessment of the system. However, nothing prevents us from using the very same approaches to prompt system adaptation and correct the system behavior as it runs.

In the end, this layer can be instantiated based on a wide range of criteria for assessment for different CONIAC systems, enabling the discovery of information in varied contexts and tasks. This is precisely the purpose of CAFE, introduced in the next section.

3.2 Evaluation Framework

Fig. 6 illustrates our evaluation framework that accounts for the diverse sets of choices of tasks, stakeholder goals, and criteria that need to be taken to evaluate CONIAC systems. We believe that our evaluation framework can not only cover most of the commonly used CONIAC evaluation schemes but can guide a system designer to develop novel evaluation schemes based on both traditional and non-traditional measurements.

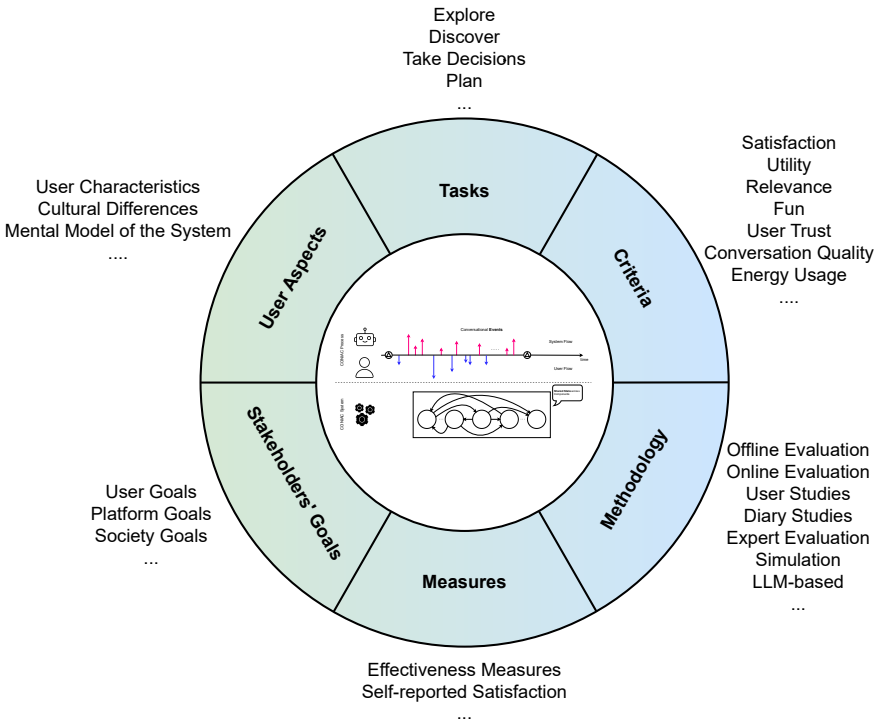

Figure 6 Evaluation Framework.

The main components of our framework account for different user aspects, stakeholder goals, tasks, criteria, evaluation methodologies, and measures. Following our evaluation framework, a system designer interested in evaluating a CONIAC system should be able to develop an evaluation scheme that can be used as a continuous or multiple-shot experimental/evaluation probe as discussed in Section 3.1.

- **Stakeholder goals.** Stakeholders of a CONIAC system may have diverse goals that might or might not be directly accessible to system designers or evaluators and must often be implicitly inferred in evaluation. CONIAC systems might also have multiple goals ranging from end users having (in-)direct information needs to platforms deploying CONIAC systems interested in content usage, user engagement, impression generation, and user retention, to name a few.
- **User aspects.** When developing an evaluation framework for CONIAC systems, it is crucial to consider user-specific aspects, such as preferences, specialized needs, expertise types, and background characteristics, which may make conversational systems more beneficial than non-conversational alternatives.
- **Tasks.** CONIAC involves tasks characterized by an information need, human involvement, goal orientation, and mixed-initiative between the user and the system. While some tasks benefit from the introduction of a conversationally competent system, others may not, depending on the complexity of the task.
- **Criteria.** The scope of evaluation can range from single-turn interactions to entire conversations and long-term system usage, each requiring different criteria for assessment. Additionally, the temporal dimension, which examines how the system's performance changes over time, is a critical factor that intersects with both stationary and dynamic properties, making it essential for a comprehensive evaluation.

- **Methodology.** In addition to the standard distinction of user-focused and system-focused methodologies, our evaluation framework categorizes evaluation methodologies also according to the employed time model – a dimension especially relevant for CONIAC. This dimension ranges from stationary methodologies like single-interaction experiments to methodologies like controlled lab studies that allow for continuous measurements, such as physiological ones.
- **Measures.** Finally, we allow for measures that typically focus on the system's ability to provide accurate, relevant, and timely information during interactions. Measures include objective measures of effectiveness to subjective notions of user satisfaction like self-reported satisfaction. By incorporating both objective effectiveness measures and subjective self-reported satisfaction, evaluators can better understand the system's strengths and areas for improvement.

We offer details for each of the items above in Section 4 to Section 9.

We also take Fig. 6 as the *workflow* to be adopted to evaluate CONIAC systems. We start from defining the "Stakeholders' Goals" and progress clock-wise to "User Aspects", "Tasks", "Criteria", "Methodology", and "Measures". All together, "Stakeholders' Goals", "User Aspects", and "Tasks" constitute the context in which the CONIAC system is evaluated, whereas "Criteria", "Methodology", and "Measures" define how the actual evaluation is to be carried out.

3.3 Limitations of Current Conversational Evaluation

The literature is rich in algorithms and strategies for handling multiple scenarios from conversation information access. While most of them are inspired by conversational search and conversational IR literature, we have seen an interest in conversational recommendations over the last few years. Going through these works, a key commonality emerges: the lack of consensus on evaluating a conversational system in a manner that gives a holistic view of its performance and enables comparing and contrasting with other emerging technologies. Indeed, TREC tracks like TREC CAsT [34, 33, 35, 89] and TREC iKAT [3, 4] offer uniformity to comparisons for conversational search, but do so primarily with the relevance-driven focus for evaluation. Datasets such as ReDial [79], ConveRSE [57], or LLM-REDIAL [81] enable assessment of conversation for recommendations. However, as we previously stated (and illustrated with the earlier example), we emphasize that CONIAC is a complex endeavor and should be evaluated from multiple lenses [134].

4 Stakeholder Goals

Stakeholders have varying goals, which are not generally accessible to system designers or evaluators and are not always explicitly formulated even for the stakeholder. They can often be inferred from the actions and the background information the system has available. Stakeholder goals are not always aligned: a system may benefit some stakeholders more than others [115, 39]. In general, stakeholders may have several simultaneous, interacting, and only partially overlapping goals [40, 12]. A conversational approach can further many stakeholder goals.

- *Users* have goals they wish to see a system help them achieve. A user may be solving an immediate problem or a long-term issue; wish to retrieve something they already know about (i.e., known item search); seek to learn or investigate a topical or thematic area [85]; they may be looking for diversion or entertainment or delight. Users may be focused on a goal, or the goal may be in the background for some other activity, and the goal may be of varying importance and centrality to the user throughout the task they are pursuing. Users may approach information access systems to find immediate factual responses, seek support for an argument, or plan a later investigation into an area. They may be seeking documents or collections of documents, they may be looking for some factual content they know or believe to be found in a document, or they may be looking to enjoy reading, viewing, or listening to some content that may be found in an item in a collection or across several items. To achieve these goals, users will generally want to expend little effort, enjoy interacting with the system, trust the correctness of the result, and feel confident that they have exhausted the usefulness of the system they are working with for the purposes they have engaged with it.
- *Indirect users* are affected by the decisions and actions of the (primary) users based on the information they gained from the system. Examples are a doctor's patients consulting medical information systems or clients of lawyers who use their legal information systems. While the primary users may experience only gradual changes in the quality of their work due to the strengths and weaknesses of the information system, their individual decisions may have serious consequences for their clients.
- *System designers* and *App designers* wish to verify that their design is appealing, appropriate, effective, efficient, and catchy.
- *Distribution platforms* wish to retain the interest and appreciation of users, creators, and advertisers.
- *Creators* provide materials they would like to see being used and enjoyed, and in many cases, they wish to receive remuneration for that usage.
- *Publishers*, aligned with creators, wish to see their catalog prominently featured among the offerings to the users.
- *Advertisers* wish to receive many impressions for their messaging and to see those impressions convert to business opportunities.
- *Editors and aggregators* wish to guide a user to further material.
- *Other stakeholders can be added here*: surrounding agents such as colleagues, friends, family, or teachers.

In the following, we mainly focused on describing user goals, not on the other stakeholders' ones. However, it should be noted here that the functions of a system must satisfy, or at a minimum address, the goals of all stakeholders to be successful.

Top-level goals

As deeply discussed in many venues [2], relevance, truthfulness, and trustworthiness[5] are primary and very common top-level goals. These top-level goals are typically well-represented and adequately considered in many evaluation activities in the field. If you look at the tasks offered by initiatives such as *Text REtrieval Conference (TREC)* [51], *Conference and Labs of the Evaluation Forum (CLEF)* [43, 42], *NII Testbeds and Community for Information access Research (NTCIR)* [100], or *Forum for Information Retrieval Evaluation (FIRE)*, you will find many evaluation efforts revolving around these top-level goals.

[5] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Minimizing effort

Users wish to minimize the effort they expend on using a system. This means the interaction must have low friction and a low threshold. A conversational system will allow users to approach the system with little preparation and low cognitive load: they do not need to formulate their needs, tasks, and goals in terms exactly acceptable to the system since this can be negotiated during the session [123]. Conversely, using a conversational system may incur overhead – instead of immediate access to a known item through a short command, a user may need to engage in a more verbose interaction. A system needs to be convenient to the frequent power user and forgiving to the novice and the occasional user. The latter has more to gain from a conversational approach.

The need for conversational systems to cater to power users and novices implies that evaluation metrics must consider varying user expertise and interaction styles, balancing efficiency and user satisfaction. Evaluators should assess systems based on their ability to minimize cognitive load and friction while providing flexible, adaptive interactions that accommodate user needs and preferences.

Establishing, clarifying, and elaborating objectives

Conversational interaction allows the system and user to disambiguate ambiguous or polysemous concepts as well as clarify and specify terminology jointly.

Conversational interaction allows systems and users to jointly explore and modify the shared understanding of some topic or theme of interest. By iterating over formulations, users can learn about the representation and terminology used in the system and of the topical coverage of a collection or a system; systems can gain a better understanding of user background and users' previous understanding of the topic or theme at hand. The ability of conversational systems to facilitate dynamic, interactive learning and exploration means that evaluation metrics must go beyond static measures of accuracy or relevance, encompassing the system's adaptability and capacity to refine user understanding over time.

Evolution and specialization of objectives

Conversational interaction allows users to incrementally build on previous exchanges naturally and intuitively. This approach facilitates the user's exploration of a topic or theme and enables the system to offer relevant suggestions and guide the user better.

This allows users to *specialize* their goals over the course of a conversation, starting with a general exploration of a topical area or a theme and then progressively drilling down into a more detailed or fine-grained view until some target of interest is found. Iterative and interactive information retrieval provides a basis to understand some of the challenges, most notably that of query reformulation, which, without the support of a system, has been known to be a demanding task for users [14].

Alternatively, this allows users find and pursue a trajectory or a curriculum through a collection to meet a less specific and possibly longer-term goal of learning or familiarizing oneself with a topic. Achieving expertise in a topic does not only involve having access to documents and materials but also acquiring skills that are imparted through conversation and apprenticeship with an expert [30, 98].

Enriching an interaction with secondary conversational goals

Among several conversational goals are those of establishing a relationship between counterparts. This is an incidental goal that in human-human interaction is addressed in parallel with informational or narrational activities [58, 25], which has been studied in human-human communication [129].

Conversational interaction allows users and systems to engage in meta-dialogue to make goals, biases, and backgrounds explicit to themselves and their counterparts. Through conversational interaction, users can assess the reliability and trustworthiness of a system through conversational moves that may not be immediately goal-directed.

Conversational interaction may add to the appeal of conversation with a system by allowing for ostensibly superficial chit-chat about the topic or theme at hand. This will allow users to gain expertise not only about the topic but also how it can be acted upon.

5 User Aspects

In this context, we refer to "users" as human agents interested in using an information access system with an *information need*. As we will discuss in Section 6, this need can be diverse, from answering questions, attaining knowledge, planning a trip, or getting recommendations. Users approach systems in the hope that the system will help them meet those needs; to this end, systems offer the user informational, educational, and entertaining materials.

Some users have preferences, special needs, or other characteristics that will make a conversational system more useful than a non-conversational one; for example, people with visual impairments, senior citizens, children, or those seeking coaching. Following principles of Universal Design [48, 88] or "Design for All", many or even most such specific preferences will lead to design solutions that improve interaction for every user in the general case. Contextual requirements may benefit from a conversational approach: hands-free operation of a system or situations where a user's visual attention is otherwise occupied are examples where a conversational system would well serve every user.

Individual differences

Users might differ in how much they would like to interact with a conversational system. For example, those with strong domain expertise might have clearer mental models of the domain and can explicate their needs efficiently. A conversational system may impose an annoying or even prohibitive overhead in such cases; expert and experienced users may want to access information through shortcuts – these can be made accessible through conversational systems. Similarly, novices would especially benefit from a conversational interaction as it might help them disambiguate their needs while learning more about the domain through the interaction, with the system being able to ask clarifying questions. It is also important to note that the same user may be a novice in one domain and an expert in another, necessitating a flexible system that can adapt to different levels of expertise across different contexts.

Users also differ in their cognitive needs around information access, e.g., in their tendency to seek information. For example, the health domain Pang et al. [91] categorizes information seekers based on their level of reading engagement and research tactics: a *knowledge digger* would be high on both, whereas a *quick fact checker* would be low on both. In decision-making, a similar distinction is made between users who want to make the best, most optimal decision possible (maximizers) versus those who just want a satisfactory outcome

Table 1 A model of health information seekers, based on Pang et al. [91].

		Reading Engagement	
		Low	**High**
Research Tactics	**Extensive**	All-around Skimmer	Knowledge Digger
	Basic	Quick Fact Seeker	Focused Reader

(satisfiers) that satisfies most of their needs. Maximizers usually engage in extensive product comparison and aspire to find the best possible option [28]. In earlier work on a critiquing recommender system, Frederix [44] found that maximizers experienced more guidance from a (conversational) critiquing system compared to a simple filtering tool, whereas novices did not. Both these examples show that conversational systems might be used, experienced, and thus evaluated quite differently depending on individual differences of the user. Using individual difference measures (such as a maximization scale) allows one to distinguish between users in the evaluation.

Cultural differences

Cultural differences in how a conversation proceeds, if not accommodated by a conversational system, may annoy and disturb users [71]. For instance, users from cultures that value directness and efficiency might prefer concise, straightforward interactions, whereas those from cultures that emphasize politeness and formality might expect more elaborate and courteous exchanges. This can influence how questions are asked, responses are given, and even the pacing of the conversation. The notion of politeness in itself is a many-faceted behavioral characteristic, and its various aspects are prioritized differently across cultures.

Moreover, cultural variations extend to preferences in communication styles, such as indirect speech, and non-verbal cues, like pauses or gestures, which a conversational system might need to interpret and adapt to. For example, in high-context cultures, much of the communication relies on implicit understanding, and users might expect the system to read between the lines rather than requiring explicit statements. Conversely, users from low-context cultures might prioritize clarity and explicitness, expecting the system to provide detailed, unambiguous responses. Language itself is also a significant cultural factor. Variations in dialects, idioms, and slang can pose challenges for a conversational system, which must recognize and appropriately respond to diverse linguistic expressions.

Failure to account for these cultural differences can lead to misunderstandings, frustration, and a decreased sense of trust and satisfaction with the system. Therefore, it is crucial for conversational systems to incorporate culturally adaptive mechanisms, such as language localization, adjustable politeness levels, and the ability to recognize and respect cultural norms and values. During the evaluation, it will also be important also to take such cultural differences into account, as there might be complex interactions between cultural differences and performance and user experience measures.

6 Task

This section explores the characteristics and tasks suitable for CONIAC systems, which are defined by factors such as information need, human involvement, goal orientation, and iterative interaction. In particular, CONIAC implies tasks that have certain characteristics, highlighted below

- an information need and knowledge gap [36]
- human-in-the-loop (i.e., involving a human user)
- goal-oriented
- iterative
- interactive and mixed initiative [55]
- task switching (multiple goals)

which we elaborate on in the following.

Traditional information retrieval and recommendation tasks share many of these characteristics, particularly the notion of an information need and a user with a goal that relies on that information. Some of these tasks are furthered by introducing a conversationally competent system, while others may not be. For example, for a "factoid" query with a specific and clear information need, it may not be beneficial for a user to use conversation to satisfy that need. However, a conversational system should still be able to respond to such queries.

In contrast, some tasks strongly benefit from the opportunities afforded by conversation, including an iterative process and agents interacting via turn-taking. Particularly, *complex* tasks [20, 22] that require clarification and refinement to support a user's goal are well-served by an interaction between the system and the user.

We propose a model for conceptualizing CONIAC tasks in Fig. 7.

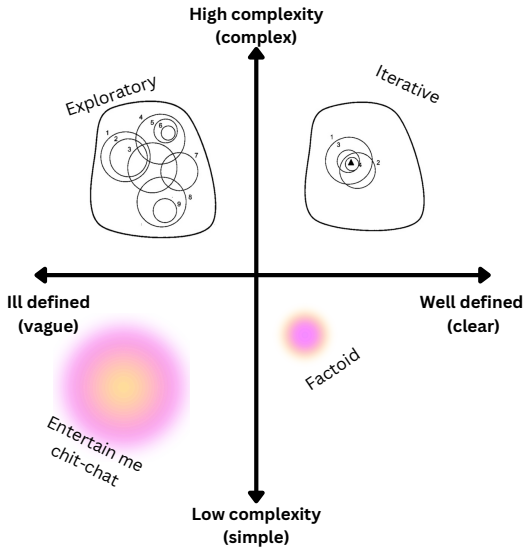

Figure 7 The tasks that are ideally solved by CONIAC systems. The model proposes two core dimensions of task complexity and task definiteness. (Iterative vs. exploratory search behavior representation adapted from White and Roth [128]).

In this model, complex tasks can vary in definiteness, with implications for the nature of an interaction between a system and a user. A complex task that is more well-defined requires primarily an iterative interaction process aimed at refining the expression of the

information need to be more focused or precise [14]. A more ill-defined information need involves exploration, a process that involves learning and investigation, and examining a broader range of information – often uncovering unanticipated information. This process typically involves refining the information need itself. Here, we draw on the notion of iterative vs. exploratory search defined by White and Roth [128]. This model explicitly reflects the interaction behaviors that are likely to be involved with these two types of search.

A user's information journey in complex tasks is well-documented in the literature [85, 76], highlighting how user states can evolve throughout the process. For instance, the Information Search Process (ISP) model [76] demonstrates that a user's initial perception of a task's complexity might start as simple but can quickly become more complex due to the broad scope of relevant information. This suggests that CONIAC systems should be capable of detecting and modeling these evolving user states, optimizing their functions and responses accordingly to enhance the user experience.

Task switching is another common characteristic of human information-seeking behavior, observed in both web search engines and conversational assistants [111, 121]. In complex tasks, users often pursue multiple sub-goals, and the outcome of each sub-goal can influence subsequent actions or significantly impact the final goal. Therefore, it is crucial for users to track multiple tasks effectively. However, traditional information access systems typically leave the burden of managing these tasks to the users.

One example of a complex but well-defined task is the task of buying a smartphone – there are myriad options, with the information spread across many sources and a large and ever-increasing number of variables to consider.

Conversation allows for both personalization and system-human negotiation to define the information need. In short, rather than a one-shot interaction between a user and a system going directly from an articulated information need to an "answer", conversation allows for a *process* to arrive at the satisfaction of the user's information need.

In the following, we illustrate typical conversation information access tasks and how they relate to dimensions of complexity and definiteness. We also illustrate how users might dynamically traverse the space during the conversational interaction with the system [117, 76].

Hypothesis generation task

The generation of hypotheses is a key component of high-level critical thinking in contexts including scientific discovery, problem-solving, and decision-making. Hypothesis generation is a complex information access task. It involves exploring an information space to identify connections between entities, variables, or events that can be tested. It is typically grounded in existing associative or relational patterns. Therefore, it lends itself well to a conversational process. Such an approach helps examine the information space for known connections and then supports a user postulating, prioritizing, and interrogating new ones.

In the context of scientific discovery, hypothesis generation involves investigating scientific literature. This often includes making connections between seemingly unrelated studies [116]. A classic example is the "ABC" literature-based discovery model introduced by Swanson [109], in which implicit shared items ("B" items) are analyzed to form links between two entities ("A", "C"). Under this model, a user could be supported by a conversational system to move logically between different parts of the information space.

Evaluation of a hypothesis generation task is challenging as there is no truly "correct" answer that the user seeks. Notions of novelty, surprise, feasibility, promisingness, or interestingness are relevant [108].

Product search and recommendation task

Suppose a user with limited domain knowledge needs to replace an old washing machine: "*I need to replace my washing machine and I want one that serves our family of five and is somewhat efficient*". We initially would classify this as a well-defined need of moderate complexity. Current online tools (e-commerce shops) would allow you to search products with some filtering and construct a set of alternatives, but it will be hard to understand the domain: e.g., what is a sufficient capacity in kg? What amount of kWh makes it efficient? Also, the number of features presented by the tool might bring the realization that there are other important things to consider (rinsing quality, time per cycle) and might overwhelm the user with the number of tradeoffs to make. Hence, a simple search might make the user aware that the need is not as well-defined and the task perhaps more complex than envisioned. However, a regular online shop will not provide the tools to help users understand the domain better. A conversational system, using techniques from critiquing recommenders [44] would help the user to understand the domain and tradeoffs better: "*based on your query, we have several efficient washing machines that have 8kg capacity or more that suit your family needs. Would you rather like a washing machine that rinses the clothes better and but will be less silent*". Based on this response, the user discovers some other (latent) needs and tradeoffs they were unaware of before, and the dialog will allow them to define them better iteratively. This will allow the user to learn about their actual preferences and the best matching products, and the system can provide a more adequate recommendation.

Travel planning task

This scenario aims to underscore the necessity for a new evaluation framework, criteria, and metrics to examine how complex conversational processes evolve between a user and a CONIAC system beyond merely focusing on the outcomes. A user approached a conversational agent to plan a family vacation. They are four people, including two children and a small dog, with a certain budget level. Initially unsure about destinations, activities, and accommodations, the user relied on the agent's guidance. The agent recommended Japan – specifically Kyoto and Tokyo – and provided detailed information supported by traveler reviews. When the user expressed concerns about keeping the children engaged, the agent asked probing questions and suggested interactive activities like the Samurai and Ninja Museum, which helped the user feel more confident in the trip's potential.

As the planning progressed, the user's uncertainty started to diminish. For flights, the agent found a well-reviewed direct option within budget, but detecting the user's hesitation about long travel times with children, the agent inquired about preferences for layovers or onboard services. This led the user to feel more assured in choosing the direct flight. Regarding accommodations, the user's initial concerns about the suitability of a traditional ryokan for their children and dog were alleviated after the agent found a more family-friendly alternative. The agent's ability to adapt recommendations based on evolving user needs gradually transformed the user's uncertainty into confidence.

By the end of the journey, as the agent reviewed the finalized itinerary, the user had moved from a state of high uncertainty to one of low uncertainty, feeling secure in their travel plans. The user confirmed the itinerary and was satisfied that the agent had addressed all concerns and tailored the trip to their family's specific needs.

Health information-seeking task

Looking for health information is a natural task for CONIAC. Many individuals already use the internet for health advice, such as using symptom checkers to determine if they might have an underlying condition or whether they should consult a healthcare professional [31].

Research has examined health information-seeking behaviors, for example, considering how users may address an open-ended task such as "Imagine you are going to a party and will discuss health information with your friends. Gather enough information for an interesting discussion." (adapted from [90]). This is a vague but complex task, requiring exploration of a large and dynamic information space, accommodation of serendipity to access novel health concepts, and interactions between the user and the system to support the identification of information that is interesting and engaging to the user.

A recent study on a CONIAC system designed to help patients find cancer-related clinical trials indicates that these systems could make health information more accessible for individuals with limited health or computer literacy skills [17].

While CONIAC has significant potential, several concerns must be addressed when implementing these systems in the health domain. One issue is that these systems may lack the necessary expertise to accurately answer all questions, which could lead to misunderstandings or misinterpretations and, consequently, incorrect responses [114]. Although this is a common challenge for all search systems, it could be particularly catastrophic with health information. Additionally, these systems often handle sensitive patient data, which requires robust safeguarding measures. Voice-only CONIAC systems might also face difficulties with speech recognition, especially when users are distressed or in noisy environments [110].

Enterprise and personal information management task

Enterprise information access and personal information access have distinct requirements compared to traditional web search engines. These include challenges such as searching across enterprise (individual) intranets or navigating multiple internal or private data sources [52]. Although there has been growing interest in workplace-oriented digital assistants, such as Alexa for Business or the Cortana Skills Kit for Enterprise, their adoption has been limited. However, with the increased adoption of LLM-based systems, new solutions like Microsoft Copilot and Google Gemini are emerging to enhance productivity and streamline information retrieval in enterprise settings [120]. Microsoft Copilot is designed to assist employees by integrating with Office applications and providing contextual assistance and insights with LLMs. Similarly, Google Gemini aims to leverage artificial intelligence to improve search and information access across Google's ecosystem. Despite these advancements, the use of these tools remains relatively low, suggesting that further development is needed to realize and understand the possible task and the system's potential in the enterprise or personal information environment fully [120, 121].

Children's information discovery in the classroom

Children regularly engage in information discovery tasks that involve exploring online resources to complete classroom assignments. Their default approach is to use web search engines, but they often face challenges: (i) difficulty in formulating their information needs into succinct keyword queries, (ii) a preference for using natural language or question-based queries, and (iii) trouble navigating *Search Engine Result Pages (SERPs)* to find relevant resources [9, 77, 95]. A conversational information access system could be highly beneficial in this context. Such a system could ask clarifying questions to help children refine their information

needs, provide specific answers when required, and recommend a limited set of resources for further exploration. However, a one-dimensional evaluation (e.g., the system's 'goodness' based on its ability to deliver the necessary information or user perception—especially, since children's perceptions of their interactions with information systems do not always align with their actions) is insufficient [19, 124]. As children are the primary stakeholders in this scenario, it is essential to consider factors like the suitability of the information provided and whether the vocabulary matches their literacy levels. This highlights the importance of explicitly considering the task at hand, the stakeholder performing the task, and the various criteria for evaluation before determining the specific metrics to be used as we reflect on the CAFE components.

Curate a background entertainment task

Let us suppose that a user starts interacting with a CONIAC system as follows: *"The next two and a half hours, I will be working on something that takes up most of my attention, gaze, and hands. I want to have something entertaining in my headphones during that time. I am interested in some general topics: personal aviation, allergy medication, history of Central Asia, and I like yacht rock and string quartets"* (the last part is presumably known to the system). The system can suggest some listening material to the user, excerpts from audiobooks and lectures, recent newscasts with local news, and playlists the user has saved in the library. The user can then respond by ranking the suggestions, helping the system to sort the items that will be played in the session. The system can verify if those preferences are generally true or specific to this session.

Longitudinal learning task

A longitudinal learning task in conversational systems refers to a sustained, adaptive learning process where the system engages with the user over an extended period, facilitating the exploration of various topics or skills. Unlike single-session interactions, longitudinal learning tasks involve multiple interactions spread over days, weeks, or months. These systems are designed to remember past conversations, track the user's progress, adapt to changing interests or knowledge levels, and personalize learning experiences based on ongoing dialogue. Longitudinal learning tasks in conversational systems can provide personalized and adaptive learning experiences by adjusting to users' evolving interests, mental models, and needs. They enhance knowledge retention through continuous engagement, leveraging past interactions (i.e., memory) for more relevant and context-aware responses. Evaluating these tasks requires a shift toward more comprehensive, long-term metrics that capture the complexity and evolving nature of sustained learning interactions.

Establishing that a system is not the right one for some task

Let us consider the following scenario. The users asks a system about interesting sights to see in Saarland and bike paths. The user gets some tips, but they do not seem to be geared toward her tastes, so the user elaborates that she wants bike paths for mountain biking and that she will be camping. The system responds but delivers the same content. The user then asks for tips on Limburg, an area she has already vacationed in and know well. The user realizes that the knowledge of this system is superficial while it was initially presented as a useful resource, and she will now turn to another system. (If not conversational, the user would have had to work hard to redo the query for another area).

Toward future tasks

We have yet to exhaustively determine the range of tasks that conversational systems will eventually be capable of performing. Recent research indicates that these potential tasks are expanding beyond traditional roles, such as *finding information* and *learning tasks*, to more complex functions like *creating*. This evolution is illustrated in Fig. 8, which shows the shift towards more creative and generative tasks. We can anticipate future developments in personalized user experiences, enhanced human-AI collaboration, advanced summarization techniques, and multilingual communication. Consequently, the methods for evaluating these evolving tasks will also need to adapt, potentially incorporating new metrics that reflect these systems' increased complexity and capability.

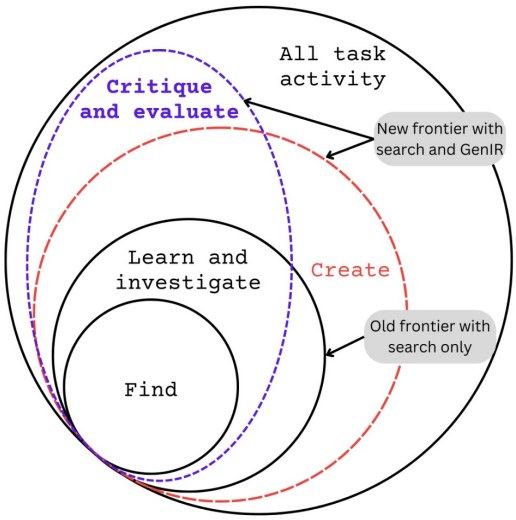

Figure 8 Visualization of possible tasks for information access, demonstrating different information access and activity levels [127, 122]. The diagram highlights the progression from simple information finding to more complex tasks like learning, investigating, critiquing, evaluating, extracting, and creating. The new frontier with Generative Information Retrieval (marked by dashes) indicates that these systems can enable advanced tasks such as critiquing and evaluating, expanding beyond the traditional search frontier also applicable to CONIAC tasks.

7 Criteria

This section discusses the criteria to consider when instantiating an evaluation framework for the conversational search scenario. As mentioned in Section 4, a comprehensive evaluation of the cooperative interaction flow between a CONIAC system and its users should involve the goals of different stakeholders. In this sense, the evaluation criteria should also account for the different perspectives on the evaluation, e.g., from a user-oriented, company-focused, or even societal perspective [60, 12, 119]. Sometimes, a stakeholder's goals and the corresponding evaluation criteria can overlap and may account for multiple aspects. However, several evaluation criteria can be clearly attributed to the different aspects of the conversational search and recommendation process, being both objective and subjective.

Fig. 9 illustrates a taxonomy of criteria that can be used to evaluate a conversational system. The first partitioning of the criteria concerns the subject of the evaluation, and it allows us to identify two major classes of evaluation criteria: *System-centric* and *User-centric* evaluation criteria. Note that the literature about criteria is rich and diversified, lacking standardized classifications for them across different research areas. It is beyond the scope of this manifesto to propose such a commonly agreed classification. Therefore, by no means do we claim the listed criteria in the taxonomy to be exhaustive, but rather, we aim to categorize the different aspects of the cooperative conversational process, providing examples of suitable evaluation criteria.

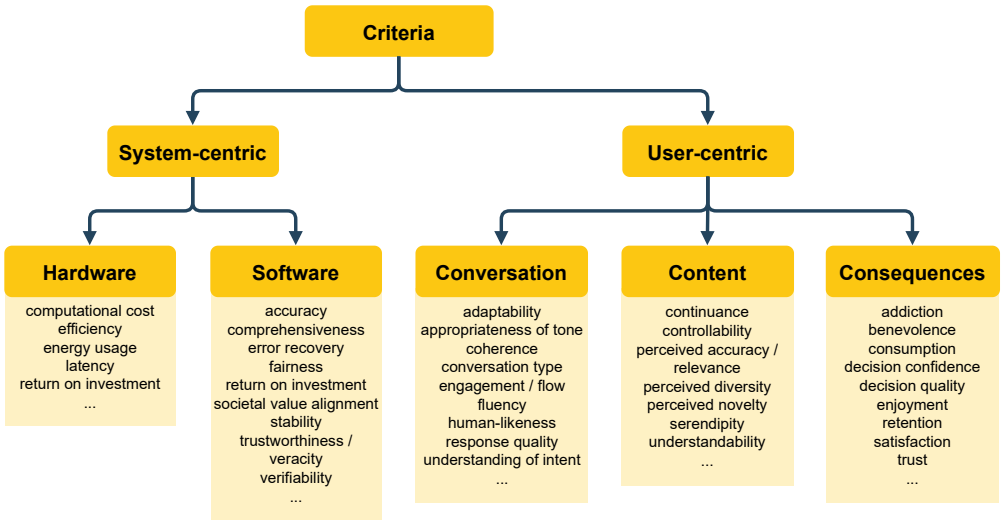

Figure 9 Taxonomy of evaluation criteria.

System-centric criteria

With system-centric evaluation criteria, we refer to those aspects typically – but not exclusively – characterized by a high level of objectivity. These criteria include those that drive the development of the system (i.e., they can be derived from the stakeholders' constraints or induced by the goals of the system) and those that can be evaluated in offline scenarios using the system in isolation, although they may require some preexisting user input, such as existing user ratings or expert labels. We can further partition system-centric criteria into *hardware* and *software* criteria.

Hardware-centric criteria

Hardware-centric evaluation criteria include, among others, the system's energy usage, computational costs, efficiency, and latency. Modern state-of-the-art CONIAC systems mostly rely on LLMs with dedicated hardware and energy requirements. As earlier work by Strubell et al. [113] pointed out, the energy usage of LLMs has a considerable environmental impact, and these implications should be kept in mind by platform operators, system developers, and also by the users. Closely related to energy usage, computational costs are a more general criterion that is of special interest to the platform operators and service providers, who eventually have to cover the resulting financial costs. Regarding the system's response time, latency and efficiency are important criteria for system operators and users alike. Long response times may dissatisfy the users, making the service less attractive, which can potentially lead to user abandonment or quitting the use of the service entirely.

Software-centric criteria

Software-centric criteria include those criteria that are not particularly bound to the hardware. Among others, the system's accuracy, comprehensiveness, or error recovery are objective criteria that can be quantified to better understand the CONIAC system's impact on the conversation when evaluating the conversational process. For instance, *accuracy* reflects how well the system provides correct responses (e.g., accurate recommendations) to the user's request. *Comprehensiveness* accounts for how thoroughly the request is answered and if it requires follow-up interactions like rephrasing or making more utterances in general. *Error recovery* refers to the quality of recognition and recovery from misunderstandings on both sides of the conversation.

Other criteria – of particular interest to the society as a whole – include fairness, replicability, and the alignment to societal values. An evaluation of fairness criteria allows for a better understanding of potential biases in the data or the algorithms. Similarly, evaluating the replicability of system-based components allows audits of the reliability and robustness of the CONIAC system, which may inherently impact other criteria like the system's trustworthiness. In this regard, the development of the system can also be driven by societal values, which, in turn, requires an evaluation of how well the CONIAC system is able to grasp and convey these values to the users in the end. In this context, *veracity* plays an important role, which accounts for the system's ability to provide responses that also reflect reality. This is especially important, considering the consequences for decision-making contexts. Furthermore, recent work by Joko et al. [63] suggests that LLM-based conversational agents are not diverse, so *diversity* can be an important criterion for future evaluations of CONIAC systems.

We refer the reader to Sakai [99], who also provides an elaborated taxonomy of criteria for evaluating textual CONIAC systems. Many of the criteria mentioned by Sakai [99] can be aligned with the different aspects we envision in our taxonomy.

User-centric criteria

User-centric criteria represent the second broader category of aspects that should be evaluated when it comes to interactive systems Kelly [66], Knijnenburg et al. [74] and especially CONIAC systems [62]. Users' subjective reflections on their interaction with the system can be characterized as their perceptions of certain qualities of the system itself and the

self-relevant consequences of their use of the system. For CONIAC systems, we divide users' perceptions of system qualities into perceptions of the **conversation** itself (e.g., engagement, response quality), and perceptions of the **content** that is recommended/retrieved by the system (e.g., perceived accuracy, serendipity).

The **consequences** of users' interaction with the system have both a behavioral and subjective component to them – the subjective component concerns users' satisfaction with the outcome of system use (e.g., decision confidence), their satisfaction with the process of using the system (e.g., fun), and their trust in the system itself. The behavioral component concerns both users' interactions within the system (e.g., consumption, retention), as well as actions that a user takes in the real world (e.g., recommending the system to friends), and the consequences thereof (e.g., decision quality).

An emphasis on temporal evaluation

It is important to note that the **temporal** dimension of the interaction is particularly important in CONIAC systems – this relates to both the *flow* of the conversation between the user and the system, the *continuance* of the cooperative interaction between the user and the system towards fulfilling the information access goal, and the development of the user's *trust* in the system [65, 10]. This means that evaluating CONIAC systems requires us to evaluate the system at various temporal scopes: evaluation can address a single turn (i.e., the traditional evaluation metrics), a conversation (e.g., flow, continuance), or the system usage across multiple conversations (e.g., trust). Measuring the interaction at all of these levels means applying methodologies and metrics that extend beyond a single interaction and that do not just cover a CONIAC systems' stationary properties but also its dynamic properties (which can be measured either continuously or at a certain discrete temporal rate).

8 Methodology

The advantage of CONIAC systems is their support for complex, rich, cooperative interactions that go beyond the back-and-forth interactions with traditional IR and RS systems. This richness of conversations must hence be reflected in the evaluation of CONIAC systems, even beyond extensions for generative systems like that of Gienapp et al. [47]. Traditional interactive IR and RS evaluations span from system-focused evaluations to user-focused evaluation methodologies [66, 134] – in the context of CONIAC systems, we advocate for an even stronger emphasis on user-focused evaluation to better capture the qualities of these complex, cooperative interactions. Additionally, the dimension of time is critical for CONIAC systems, which is not reflected in most IR and RS evaluation methodologies used so far.

Therefore, we categorize evaluation methodologies – both quantitative and qualitative – along two dimensions: (1) according to the focus of a study (a standard dimension in IR), ranging from system-focused methodologies like offline simulations to user-focused methodologies like qualitative interviews; and (2) according to the employed time model (a dimension from system behavior theory that is especially suitable for CONIAC studies), ranging from stationary methodologies like single-interaction experiments to methodologies like longitudinal user studies that allow for continuous measurements of, for example, satisfaction. While evaluation criteria are defined as agnostic of time models, for the actual evaluation, one has to decide which time model to use, which limits the available evaluation

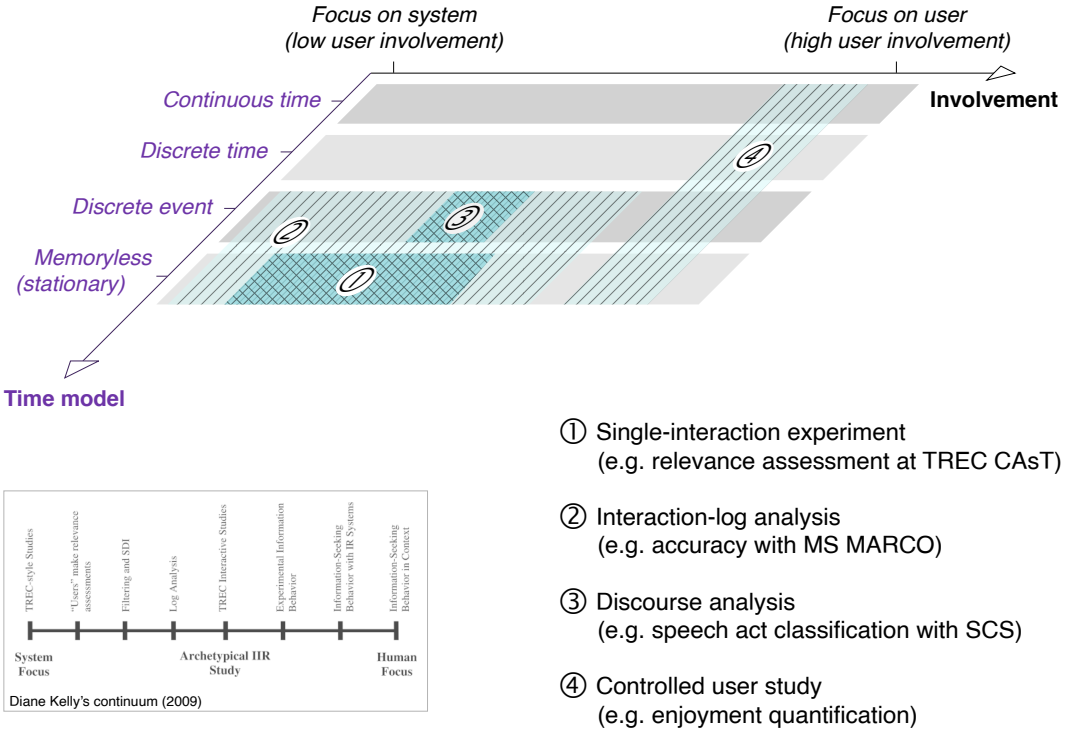

Figure 10 In conversational settings, time makes the difference. See above the methodologies spectrum, "Involvement" (the x-axis) as introduced by Kelly [66] (shown lower left), extended by an orthogonal "Time model" axis. The figure also gives four examples of methodologies and how they cover different parts of the user involvement and time model space.

methodologies and affects the details of their implementation. Given the criteria of interest, one has to choose an evaluation methodology that covers these criteria. In this respect, we give a brief overview of existing methodologies, highlighting the clear divide between user-centric and system-centric approaches.

Methodologies for CONIAC systems

Evaluation methodologies for information systems can be placed along a continuum [66] from focusing on the system (without integrating actual users, like offline simulation) to focusing on the users (without integrating specific systems, like behavioral diary studies [118]). Methodologies focusing on the interaction between specific users and specific systems are placed in the center. The space depicted in Fig. 10 has this continuum as its horizontal axis.

Conversations are complex, dynamic interactions that evolve continuously as they progress toward the user's goal. This temporal aspect corresponds to two different concepts: The *scope* (how much we measure, from single turns to multi-conversation interactions) and the *model of time* (unit of analysis, from single discrete to continuous). We expand the one-dimensional continuum of Kelly [66] with the time model that the methodology implements, taken from the theory of system behavior [132] (cf. Fig. 10). Therefore, evaluation methodologies must be selected or adapted to fit the concept of time the research question demands. The four different time models, each implemented by several methodologies, are the following:

Memoryless (stationary)

A single data point is measured per conversation (named "memoryless" in system behavior theory). For some research questions, especially when comparing systems, it is sufficient to measure either only once for a conversation – typically at the end – or to take only an aggregated measurement like an average. A typical example of system-focused evaluations with this time model are single-interaction experiments, which include the TREC CAsT [34, 33, 35, 89] and TREC iKAT [3] competitions in which systems are asked to respond to a user utterance with conversation history, or the Netflix prize in which algorithms are asked to increase recommendation performance across a dataset [16]. A typical example of user-focused evaluations with this time model are controlled user studies with a post-conversation survey [72].

Discrete event

The occurrence of discrete events is measured. Typically employed events are the utterances in conversations with chat-based agents and speaker changes for voice-based agents. Other suitable resolutions are single words uttered, topic switches, when the user has completed some sub-tasks and asks the agent for new instructions (e.g., when the agent assists in cooking), or when the recommended song is finished. A typical example of evaluation methodology is interaction-log analysis, where the events are reconstructed from the logs. Examples of available search interaction-logs include MS MARCO [87] Conversational Search[6] and CIRQL [121]. Detailed behavioral logs may be created by employing advanced user tracking mechanisms [130]. If no logs of real interactions are available, offline simulation can be employed to generate interactions, for example, using software like SimIIR [136], GenIRSim [69], and UserSimCRS [1].

One evaluation methodology that follows a discrete event time model that is especially relevant to CONIAC is discourse analysis, which entails parsing the conversation according to theories from the discourse and dialog literature as an intermediary step for the evaluation. The parsed representation of the conversation follows, for example, rhetorical structure theory [84], models of argumentation [112], or the conversational roles model [107] – the last was specifically developed for CONIAC and builds on top of the theory of speech acts [103]. The parsed representation can then be employed as a model of the conversation [70], for example, to detect anomalous conversations [125]. Available datasets include SCS [123] and MANtIS [94].

Discrete time

Measurements are taken at discrete and regular points of time according to a selected sampling rate, ranging from a fraction of seconds to days. As conversations are typically analyzed in terms of the events described above, this time model is not as frequent in CONIAC experiments. However, especially for wearable or ubiquitous conversational agents like ones integrated into smart watches or glasses, experience sampling [92] – which uses the discrete time model by definition – could be an especially suitable methodology to gain insights into natural usage behavior. Discrete-time evaluations may also be used in longitudinal studies to track the evolution of users' interactions with the system. Finally, in descriptive research, diary studies can be used to obtain user input at set time intervals, e.g., on a daily basis [24].

[6] An artificial log generated based on real search engine usage
`https://microsoft.github.io/MSMARCO-Conversational-Search/`

Continuous time

Measurements are taken continuously. Though measurements taken with computers might not be genuinely continuous, methodologies that use this time model could use a very high sampling rate to effectively resemble continuous measurements and thus allow for calculating measurement derivatives. Physiological measurements are often taken continuously, for example, saccades and dwell times in eye tracking studies [13]. Continuous measurements do not need to be without interruption. When evaluating several interaction sessions with CONIAC systems, the measurement can pause.

Existing methodologies

In the following, we briefly define well-known methodologies that can be used to evaluate CONIAC systems, some of which have been mentioned above.

Single-interaction experiment. The class of offline methodologies that examine a single user interaction with a conversational system. This technique is among the easiest methodologies for rapid assessment.

Offline simulation (incl. A/B). Offline simulation involves testing a system or model using pre-recorded data instead of real-time data [56]. Offline A/B testing compares two versions of a system by running simulations on historical data to see which performs better. These methods help in evaluating changes without affecting live users.

User simulation (incl. LLM-based). User simulation refers to methods where user behaviors are simulated by software. Especially, LLM-based approaches allow for simulating users' interaction behaviors for various tasks such as conversational search and recommendation [104, 137].

Online A/B testing. Online A/B testing involves comparing two versions of a system to see which performs better [75]. Users are randomly shown either version A or version B, and their interactions are tracked to measure effectiveness and/or efficiency.

Interaction-log analysis. Interaction-log analysis involves studying past user interactions to understand how people use a system. By examining these logs, patterns, and trends can be identified in user behavior. This information helps improve the conversational system to meet user preferences better.

Expert evaluation. Expert (or editorial) evaluation involves having specialists or experienced editors assess components of conversations. The experts provide feedback based on their knowledge of the field.

Discourse analysis. Discourse analysis is a research field for the offline study of transcripts from written or spoken conversations. Discourse analysis can be done on various levels of abstraction, ranging from the very fine-grained *conversation analysis* family of methods [106] to methods based on, e.g., rhetorical structures, social processes, or systemic linguistics. In general, discourse analysis examines how linguistic and, in some cases, para-linguistic mechanisms and elements are used to achieve communicative goals in recorded interactions and how they make it possible for participants to make sense of a conversation [49].

Data donation. Data donation methodologies involve different ways people can share their data for research or analysis [135]. This can include directly providing data, allowing access to existing data, or contributing anonymized data collected by devices. These methods help researchers gather valuable information while respecting privacy and consent. They also allow the gathering of information that sits across multiple systems or authorities.

Table 2 Examples of methods suitable for evaluating criteria categories (✓– all criteria of this category, * – only some)

Methodology	Criteria				
	System-centric		User-centric		
	Hardware	Software	Conversation	Content	Consequences
single-interaction experiment	✓	*			
offline simulation (incl. A/B)	✓	✓			
user simulation (incl. LLM-based)	✓	✓			
online A/B testing	*	✓			*
interaction-log analysis	✓	*	*		*
expert evaluation			*	*	
discourse analysis			✓	*	✓
data donation					*
wizard of oz / observational study			✓	*	*
controlled user study (incl. longitudinal)	*	*	✓	✓	✓
user survey (quantitative)			✓	✓	*
user interview (qualitative)			✓	✓	*
diary study / experience sampling			✓	✓	✓

Wizard of Oz / observational study. The Wizard of Oz methodology involves creating a prototype where users interact with a system they believe to be automated but can be operated by a human behind the screen [32, 73]. This enables researchers to test and refine the system's design and functionality before fully developing the technology. Alternatively, for conversational experiments, two humans could observe their interactions as a first step, referred to as observational studies [123]. It is a cost-effective, rapid deployment way to gather user feedback and improve the system based on real interactions.

Controlled user study (incl. longitudinal). Controlled user studies involve observing and testing users in a structured environment to evaluate how they interact with a system [72]. Longitudinal studies extend this by following the same users (commonly referred to as panels) over a longer period to see how their interactions and experiences change over time. Commonly explicit feedback is gathered after the study. These methods help researchers understand both immediate and long-term effects.

User survey (quantitative). User survey methodologies involve collecting feedback from people through questionnaires. These surveys ask users to outline their experiences, preferences, and opinions to gather valuable insights.

User interview (qualitative). Qualitative user interview methodologies involve conducting in-depth conversations with users to understand their experiences, thoughts, and feelings. These interviews typically follow a methodology (e.g., grounded theory [27]). This approach helps researchers gain an understanding of motivations.

Diary study / experience sampling. Diary study and experience sampling methodologies involve participants reporting their thoughts, feelings, and activities. In the former, participants write diary entries. In the latter, participants are prompted throughout the day. This approach helps researchers understand people's experiences and behaviors in their natural environments.

For planning an evaluation, one usually starts from the goals of the evaluation (see Section 4) and then defines the criteria of interest (see Section 7). In the next step, one needs to choose an evaluation methodology covering these criteria. Table 2 shows the methodologies described above and lists the criteria categories of Section 7 for which they are most suitable. As the table shows, there is a mostly clear divide between system- and user-oriented methodologies [66].

9 Measures

This section presents the commonly used measures for assessing system-centric hardware and software criteria, and the measures for evaluating user-centric criteria (including users' subjective reflection on their interactions, their perceptions, and behavioral measures). Depending on the goals and criteria a specific system aims to fulfill, one can adopt the suitable evaluation methodology (see examples in Table 2) to assess the corresponding measures.

Measures for system-centric criteria

There are tutorials and reviews of measures in IR [101], RS [50], and software [41]. As with all measures, it is important to ensure they are used with care [45].

Measures are commonly used to quantify criteria. For hardware, common measures include computational cost, potentially specified in currency units; efficiency, potentially specified in terms of computational needs per data unit processed; energy usage, potentially specified in kilowatt hours per computational task; and latency, potentially specified in seconds.

Software measures include, but are not limited, to accuracy, potentially specified in *Normalized Discounted Cumulated Gain (nDCG)* or *Area Under the ROC Curve (AUC)*; comprehensiveness, potentially specified in terms of recall; error recovery, potentially specified in terms of mean time to repair; fairness, potentially specified in terms of the divergence from expected distribution; replicability, potentially specified in terms of the ease with which others can reproduce an existing algorithm/system; stability, potentially specified in terms of mean time between failure; veracity, potentially specified by data error rate; and verifiability, potentially specified by the percentage of fact indication covered by evidentiary sources.

Measures for user-centric criteria

Measuring the effects of a CONIAC system on its users requires either observing the user's interaction with the system or probing the user for their reflection on their interaction with the system.

Users' subjective reflection on their interaction with the system can be measured using multi-item measurement scales, many of which can be captured by existing user-centric evaluation frameworks Jin et al. [62], Knijnenburg et al. [74] or established scales for specific aspects (e.g., for various dimensions of trust [86]. It is important to adapt these scales to the context of the system (e.g., a scale for goal achievement should broadly refer to the goal of the specific system under evaluation, if known) or to develop new scales where needed – the interested reader can refer to DeVellis [37] for a primer on how to develop robust subjective measurement scales.

Behavioral measures that comprise users' interactions with the system can be measured using conversation logs and compiled into measures that have been defined extensively in the area of A/B testing [75]. Objective outcomes that occur outside the system have been covered in the domain of behavioral psychology [131].

It is important to note that users' perceptions and behaviors (both within and outside the system) are subject to be affected by users' personal characteristics and contextual variables [21]. It is also important to measure users' perceptions and behaviors at the

appropriate temporal scale and/or frequency. For example, users' decision-making can be traced temporally through expansive interaction logging [102], and the development of trust within an interaction session can be measured by repeatedly probing users with trust-related questions or by repeatedly observing trust-related behaviors [64, 65].

10 Research Directions

In this section, we highlight some possible questions that remain open and may constitute future research directions, requiring further and deeper investigation.

Stakeholder Goals

- What do users really expect from CONIAC systems?
- How can CONIAC systems accurately deduce stakeholders' goals from interactions with users?
- How can we help users select the best CONIAC for a given task?

User Aspects

- Which types of users can benefit substantially from CONIAC systems and how?
- What are good measures of individual and cultural differences for users of CONIAC systems?
- What individual and cultural differences are important moderators of the success of (specific types of) CONIAC systems?
- How do we ensure statistical validity and power when experimenting with different types of users?

Task

- What common characteristics do complex tasks have that can benefit significantly from CONIAC systems?
- How can CONIAC systems become closely integrated into task environments, and how can their contribution to task performance be evaluated in these settings?
- What temporal patterns can be observed in the level of perceived task complexity during interactions with a CONIAC system?
- How can a user and a CONIAC system efficiently agree that a given task is beyond the system's capabilities?
- What array of criteria and methods can capture, account for, and reflect the dynamic, evolving, and possibly multi-modal nature of conversations?

Criteria

- What criterion makes a difference in users' evaluation of CONIAC systems?
- Can we foresee a future when we will be able to assess user-centric criteria (e.g., during system development) using LLMs in a simulation setting?
- How can we structurally report on multi-criterion evaluations that include both system-centric and user-centric criteria?

Methodology

- What methods can be developed to evaluate CONIAC systems that prevent test data leakage while ensuring robust and reliable evaluation metrics?
- How can we accurately attribute the performance of a conversational model to its individual components?
- What approaches can be used to evaluate the accuracy and relevance of responses from a CONIAC system, particularly when these responses are derived from the aggregation of multiple documents and the system's internal knowledge?
- Can we reuse interactions and corresponding labels between a system and a user to evaluate a different system, or would this bias our observations on the second system?
- How do we ensure reproducibility of experiments when they involve users, interaction, possibly simulation, and inherently stochastic systems (e.g., LLMs)?

Measures

- Besides user studies, what offline evaluation measures can we develop to assess continuous criteria such as conversation flow and continuance?
- What is the most effective and statistically sound approach to aggregate measures across different interactions? Is it just averaging, or should we develop protocols more aligned with the structure of a conversation (e.g., non-independent interactions, graph structure)?
- Which measures are suited to evaluate the dynamic evolution of a conversation over time, including changes in mental models and user-centric evaluation criteria?

11 Conclusions

In this manifesto, we made a case for the need for more advanced conversational systems (Section 1), which go beyond what current technology – despite being already very performing – is able to do. In this respect, we elaborated an abstract and essential *World Model* (Section 2) for the area we called *CONversational Information ACcess (CONIAC)*. In this model, we defined how a conversation goes on as a sequence of events in the CONIAC process layer and introduced the main features of CONIAC systems in terms of: (i) blending different technological domains, such as IR and RS; (ii) embedding both internal knowledge about the user and external knowledge about the world or context in which the user and the system operate; (iii) tracking the state of the conversation both as flows of events and as dynamically updated beliefs in the user information state.

The above vision of the CONIAC world led us to elaborate the *Conversational Agents Framework for Evaluation (CAFE)* (Section 3). This stems from the observation that, to align with the flows of the events in the CONIAC process layer and with the evolution of the state of a CONIAC system, evaluation should turn into a dynamic series of probes rather than being a single static assessment, as it is mostly today.

Then, in the following sections, we detailed the main areas of the CAFE framework, namely, Stakeholder Goals (Section 4), User Aspects (Section 5), Task (Section 6), Criteria (Section 7), Methodology (Section 8), and Measures (Section 9). We provided suggestions and examples of possible instantiations for each of these areas without pretending to be exhaustive.

Finally, we presented open research questions that have emerged during the discussions on the different parts of CAFE that may suggest future research directions.

This manifesto should not only be regarded as a useful account of an important research challenge. We hope that it will also produce valuable fall-outs, such as bringing these issues into the research agenda of the involved communities, providing a common ground to coherently develop cross-sectorial research, helping funding agencies envision appropriate funding instruments for addressing these challenges, and motivating researchers to overcome today's limitations.

12 Participants

- Avishek Anand
TU Delft – Delft, NL
- Christine Bauer
University of Salzburg – Salzburg, AT
- Timo Breuer
TH Köln, DE – Köln, DE
- Li Chen
Hong Kong Baptist University – Hong Kong, China
- Guglielmo Faggioli
University of Padua – Padova, IT
- Nicola Ferro
University of Padua – Padova, IT
- Ophir Frieder
Georgetwon University – Washington, DC, USA
- Norbert Fuhr
Universität Duisburg-Essen – Duisburg, DE
- Hideo Joho
University of Tsukuba – Ibaraki, JP
- Jussi Karlgren
Silo AI – Helsinki, FI
- Johannes Kiesel
Bauhaus-Universität Weimar – Weimar, DE
- Bart Knijnenburg
Clemson University – Clemson, SC, USA
- Lien Michiels
imec-SMIT, Vrije Universiteit Brussel & University of Antwerp – Brussels & Antwerp, BE
- Andrea Papenmeier
University of Twente – Twente, NL
- Maria Soledad Pera
TU Delft – Delft, NL
- Aldo Lipani
University College London – London, UK
- Mark Sanderson
RMIT University – Melbourne, AU
- Scott Sanner
University of Toronto – Toronto, CA
- Benno Stein
Bauhaus-Universität Weimar – Weimar, DE
- Johanne Trippas
RMIT University – Melbourne, AU
- Karin Verspoor
RMIT University – Melbourne, AU
- Martijn C. Willemsen
TU Eindhoven & JADS – Eindhoven & 's-Hertogenbosch, NL

References

1 Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. Usersimcrs: A user simulation toolkit for evaluating conversational recommender systems. In Tat-Seng Chua, Hady W. Lauw, Luo Si, Evimaria Terzi, and Panayiotis Tsaparas, editors, *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM 2023, pages 1160–1163. Association for Computing Machinery, 2023. doi:10.1145/3539597.3573029.

2 AI HLEG. Ethics guidelines for trustworthy AI, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

3 Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. TREC iKAT 2023: The interactive knowledge assistance track overview. *CoRR*, abs/2401.01330, 2024. doi:10.48550/arXiv.2401.01330.

4 Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. TREC iKAT 2023: A test collection for evaluating conversational and interactive knowledge assistants. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, pages 819–829. Association for Computing Machinery, 2024. doi:10.1145/3626772.3657860.

5 James Allan, Eunsol Choi, Daniel P. Lopresti, and Hamed Zamani. Future of Information Retrieval Research in the Age of Generative AI. Technical report, Computing Research Association (CRA), Washington, DC, USA, December 2024. CCC Workshop Report. doi:10.48550/arXiv.2412.02043.

6 Garrett Allen, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. Multi-perspective learning to rank to support children's information seeking in the classroom. In *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT 2023, pages 311–317. IEEE, 2023. doi:10.1109/WI-IAT59888.2023.00050.

7 Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports*, 9(11):34–83, 2020. doi:10.4230/DagRep.9.11.34.

8 Avishek Anand, Venktesh V, Abhijit Anand, and Vinay Setty. Query understanding in the age of large language models, 2023. doi:10.48550/arXiv.2306.16004.

9 Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera, and Jerry Alan Fails. Online searching and learning: YUM and other search tools for children and teachers. *Information Retrieval Journal*, 20(5):524–545, 2017. doi:10.1007/s10791-017-9310-1.

10 Leif Azzopardi, Charles L. A. Clarke, Paul B. Kantor, Bhaskar Mitra, Johanne R. Trippas, Zhaochun Ren, Mohammad Aliannejadi, Negar Arabzadeh, Raman Chandrasekar, Maarten de Rijke, Panagiotis Eustratiadis, William R. Hersh, Jin Huang, Evangelos Kanoulas, Jasmin Kareem, Yongkang Li, Simon Lupart, Kidist Amde Mekonnen, Adam Roegiest, Ian Soboroff, Fabrizio Silvestri, Suzan Verberne, David Vos, Eugene Yang, and Yuyue Zhao. Report on the search futures workshop at ECIR 2024. *SIGIR Forum*, 58(1):1–41, 2024. doi:10.1145/3687273.3687288.

11 Krisztian Balog. Conversational ai from an information retrieval perspective: Remaining challenges and a case for user simulation. In *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems*. CEUR, 2021. URL: https://ceur-ws.org/Vol-2950/paper-03.pdf.

12 Christine Bauer and Eva Zangerle. Leveraging multi-method evaluation for multi-stakeholder settings. In Oren Sar Shalom, Dietmar Jannach, and Ido Guy, editors, *1st Workshop on the Impact of Recommender Systems, co-located with 13th ACM Conference on Recommender Systems (ACM RecSys 2019)*, volume 2462 of *ImpactRS '19*, 2019. URL: https://ceur-ws.org/Vol-2462/short3.pdf.

13 Thomas Beckers and Dennis Korbar. Using eye-tracking for the evaluation of interactive information retrieval. In Shlomo Geva, Jaap Kamps, Ralf Schenkel, and Andrew Trotman, editors, *Comparative Evaluation of Focused Retrieval - 9th International Workshop of the Inititative for the Evaluation of XML Retrieval (INEX 2010)*, volume 6932 of *Lecture Notes in Computer Science*, pages 236–240. Springer, 2010. doi:10.1007/978-3-642-23577-1_21.

14 Nicholas J. Belkin, Colleen Cool, Diane Kelly, S-J Lin, SY Park, Jose Perez-Carballo, and Cynthia Sikora. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management*, 37(3):403–434, 2001. doi:10.1016/S0306-4573(00)00055-8.

15 Nicolas J. Belkin. Helping people find what they don't know. *Communications of the ACM*, 43(8):58–61, August 2000. doi:10.1145/345124.345143.

16 James Bennett and Stan Lanning. The Netflix Prize. In *In KDD Cup and Workshop in conjunction with KDD*, 2007.

17 Timothy W Bickmore, Dina Utami, Robin Matsuyama, and Michael K Paasche-Orlow. Improving access to online health information with conversational agents: A randomized controlled experiment. *Journal of Medical Internet Research*, 18(1):e1, January 2016. doi:10.2196/jmir.5239.

18 Craig Boutilier. A POMDP formulation of preference elicitation problems. In *Eighteenth National Conference on Artificial Intelligence*, pages 239–246. American Association for Artificial Intelligence, 2002. doi:10.5555/777092.777132.

19 Robin Burke, Gediminas Adomavicius, Toine Bogers, Tommaso Di Noia, Dominik Kowald, Julia Neidhardt, Özlem Özgöbek, Maria Soledad Pera, Nava Tintarev, and Jürgen Ziegler. De-centering the (traditional) user: Multistakeholder evaluation of recommender systems. *Int. J. Hum. Comput. Stud.*, 203:103560, 2025. doi:10.1016/J.IJHCS.2025.103560.

20 Katriina Byström and Kalervo Järvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, 1995. doi:10.1016/0306-4573(95)80035-R.

21 Wanling Cai, Yucheng Jin, and Li Chen. Impacts of personal characteristics on user trust in conversational recommender systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3491102.3517471.

22 Donald J. Campbell. Task complexity: A review and analysis. *Academy of Management Review*, 13(1):40–52, 1988. doi:10.5465/amr.1988.4306775.

23 Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1), 2012. doi:10.1145/2071389.2071390.

24 Scott Carter and Jennifer Mankoff. When participants do the capturing: the role of media in diary studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 899–908, New York, NY, USA, 2005. Association for Computing Machinery. doi:10.1145/1054972.1055098.

25 John P Caughlin. A multiple goals theory of personal relationships: Conceptual integration and program overview. *Journal of Social and Personal Relationships*, 27(6):824–848, 2010. doi:10.1177/0265407510373262.

26 Yi Chang and Hongbo Deng. *Query understanding for search engines*. Springer, 2020. doi:10.1007/978-3-030-58334-7.

27 Kathy Charmaz. Grounded theory. In Jonathan A. Smith, editor, *Qualitative psychology: A practical guide to research methods*, volume 3, pages 53–84. SAGE, 2015.

28 Nathan N. Cheek and Barry Schwartz. On the meaning and measurement of maximization. *Judgment and Decision Making*, 11(2):126–146, 2016. doi:10.1017/S1930297500007257.

29 Edgar F. Codd. Seven steps to rendezvous with the casual user. In J. W. Klimbie and K. L. Koffeman, editors, *Proceedings of the IFIP Working Conference Data Base Management*, pages 179–200. North-Holland, 1974.

30 Allan Collins, John Seely Brown, and Susan E Newman. Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In *Knowing, learning, and instruction*, pages 453–494. Routledge, 2018. doi:10.4324/9781315044408-14.

31 Sebastian Cross, Ahmed Mourad, Guido Zuccon, and Bevan Koopman. Search engines vs. symptom checkers: A comparison of their effectiveness for online health advice. In *Proceedings of the Web Conference 2021*, WWW '21, pages 206–216, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3442381.3450140.

32 Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of Oz studies: why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces*, IUI '93, pages 193–200, New York, NY, USA, 1993. Association for Computing Machinery. doi:10.1145/169891.169968.

33 J. Dalton, C. Xiong, and J. Callan. CAsT 2020: The Conversational Assistance Track Overview. In Ellen M. Voorhees and A. Ellis, editors, *The Twenty-Ninth Text REtrieval Conference Proceedings (TREC 2020)*. National Institute of Standards and Technology (NIST), Special Publication 1266, Washington, USA, 2021.

34 Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. TREC CAsT 2019: The conversational assistance track overview, 2020. arXiv:2003.13624.

35 Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. TREC CAsT 2021: The conversational assistance track overview. In Ian Soboroff and Angela Ellis, editors, *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2021. URL: https://trec.nist.gov/pubs/trec30/papers/Overview-CAsT.pdf.

36 Brenda Dervin. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In Jack D. Glazier and Ronald R. Powell, editors, *Qualitative research in information management*, pages 61–84. Libraries Unlimited, 1992.

37 Robert F DeVellis. *Scale development: Theory and applications*. SAGE, Thousand Oaks, Calif., 2011.

38 Tommaso Di Noia, Guglielmo Faggioli, Marco Ferrante, Nicola Ferro, Fedelucio Narducci, Raffaele Perego, and Giuseppe Santucci. CAMEO: Fostering joint conversational search and recommendation. In M. Atzori, P. Ciaccia, M. Ceci, F. Mandreoli, D. Malerba, M. Sanguinetti, A. Pellicani, and F. Motta, editors, *Proc. 32nd Italian Symposium on Advanced Database Systems (SEBD 2024)*, pages 290–301. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, https://ceur-ws.org/Vol-3741/, 2024. URL: https://ceur-ws.org/Vol-3741/paper33.pdf.

39 Michael D. Ekstrand, Lex Beattie, Maria Soledad Pera, and Henriette Cramer. Not just algorithms: Strategically addressing consumer impacts in information retrieval. In *European Conference on Information Retrieval*, pages 314–335, Berlin, Heidelberg, 2024. Springer. doi:10.1007/978-3-031-56066-8_25.

40 Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1-2):1–177, 2022. doi:10.1561/1500000079.

41 Norman E. Fenton and James Bieman. *Software Metrics: A Rigorous & Practical Approach*. Chapman and Hall/CRC, USA, 3rd edition, 2014.

42 Nicola Ferro. What Happened in CLEF... For Another While? In L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024) – Part I*, pages 3–57. Lecture Notes in Computer Science (LNCS) 14958, Springer, Heidelberg, Germany, 2024. doi:10.1007/978-3-031-71736-9_1.

43 Nicola Ferro and Carol Peters, editors. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2019.

44 Anne P. J. Frederix. Investigating the effect of using personalized critiques in a decision-making tool. Master's thesis, Eindhoven University of Technology, 2021. URL: https://pure.tue.nl/ws/files/174846002/Master_Thesis_A.P.J._Frederix_1396676.pdf.

45 Norbert Fuhr. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum*, 51(3):32–41, December 2017. doi:10.1145/3190580.3190586.

46 Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval (FnTIR)*, 13(2–3):27–298, February 2019. doi:10.1561/1500000074.

47 Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. Evaluating Generative Ad Hoc Information Retrieval. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, pages 1916–1929. Association for Computing Machinery, July 2024. doi:10.1145/3626772.3657849.

48 Selwyn Goldsmith. *Universal design*. Routledge, 2007.

49 Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

50 Asela Gunawardana, Guy Shani, and Sivan Yogev. Evaluating recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 547–601. Springer US, New York, NY, 2022. doi:10.1007/978-1-0716-2197-4_15.

51 Donna K. Harman and Ellen M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (MA), USA, 2005.

52 David Hawking. Challenges in enterprise search. In *Proceedings of the 15th Australasian Database Conference - Volume 27*, ADC '04, pages 15–24, AUS, 2004. Australian Computer Society, Inc. URL: http://crpit.scem.westernsydney.edu.au/abstracts/CRPITV27Hawking1.html.

53 Jerry R Hobbs. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979. doi:10.1207/s15516709cog0301_4.

54 Jerry R Hobbs. Towards an understanding of coherence in discourse. In *Strategies for natural language processing*, pages 223–243. Psychology Press, 2014.

55 Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 159–166. Association for Computing Machinery, 1999. doi:10.1145/302979.303030.

56 Chih-Wei Hsu, Martin Mladenov, Ofer Meshi, James Pine, Hubert Pham, Shane Li, Xujian Liang, Anton Polishko, Li Yang, Ben Scheetz, and Craig Boutilier. Minimizing live experiments in recommender systems: User simulation to evaluate preference elicitation policies. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2925–2929. Association for Computing Machinery, 2024. doi:10.1145/3626772.3661358.

57 Andrea Iovine, Fedelucio Narducci, and Marco de Gemmis. A dataset of real dialogues for conversational recommender systems. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2481/paper37.pdf.

58 Chris Segrin James Price Dillard and Janie M. Harden. Primary and secondary goals in the production of interpersonal influence messages. *Communication Monographs*, 56(1):19–38, 1989. doi:10.1080/03637758909390247.

59 Dietmar Jannach. Evaluating conversational recommender systems: A landscape of research. *Artificial Intelligence Review*, 56(3):2365–2400, 2023. doi:10.1007/s10462-022-10229-x.

60 Dietmar Jannach and Christine Bauer. Escaping the McNamara Fallacy: toward more impactful recommender systems research. *AI Magazine*, 41(4):79–95, 2020. doi:10.1609/aimag.v41i4.5312.

61 Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Computing Surveys*, 54(5):105:1–36, 2022. doi:10.1145/3453154.

62 Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. Crs-que: A user-centric evaluation framework for conversational recommender systems. *ACM Transactions on Recommender Systems*, 2(1), March 2024. doi:10.1145/3631534.

63 Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. Doing personal LAPS: llm-augmented dialogue construction for personalized multi-session conversational search. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International*

ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pages 796–806. Association for Computing Machinery, 2024. doi:10.1145/3626772.3657815.

64 Patricia K. Kahr, Gerrit Rooks, Chris Snijders, and Martijn C. Willemsen. The trust recovery journey: The effect of timing of errors on the willingness to follow AI advice. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 609–622, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3640543.3645167.

65 Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. Understanding trust and reliance development in ai advice: Assessing model accuracy, model explanations, and experiences from previous interactions. *ACM Transactions on Interactive Intelligent Systems*, 14(4), December 2024. doi:10.1145/3686164.

66 Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2):1–224, 2009. doi:10.1561/1500000012.

67 Kimiya Keyvan and Jimmy Xiangji Huang. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *ACM Computing Surveys*, 55(6):105:1–36, June 2022. doi:10.1145/3534965.

68 Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. Clarifying false memories in voice-based search. In Leif Azzopardi, Martin Halvey, Ian Ruthven, Hideo Joho, Vanessa Murdock, and Pernilla Qvarfordt, editors, *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*, pages 331–335. ACM, 2019. doi:10.1145/3295750.3298961.

69 Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. Who Will Evaluate the Evaluators? Exploring the Gen-IR User Simulation Space. In Lorraine Goeuriot, Philippe Mulhem, Georges Quénot, Didier Schwab, Giorgio Maria Di Nunzio, Laure Soulier, Petra Galuscakova, Alba Garcia Seco Herrera, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, September 2024. Springer. doi:10.1007/978-3-031-71736-9_11.

70 Johannes Kiesel, Lars Meyer, Martin Potthast, and Benno Stein. Meta-information in conversational search. *ACM Transactions on Information Systems*, 39(4), August 2021. doi:10.1145/3468868.

71 Min-Sun Kim. Cross-cultural comparisons of the perceived importance of conversational constraints. *Human Communication Research*, 21(1):128–151, 1994. doi:10.1111/j.1468-2958.1994.tb00343.x.

72 Bart P. Knijnenburg and Martijn C. Willemsen. Evaluating Recommender Systems with User Experiments. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 309–352. Springer US, 2015. doi:10.1007/978-1-4899-7637-6_9.

73 Bart P. Knijnenburg and Martijn C. Willemsen. Inferring capabilities of intelligent agents from their external traits. *ACM Transactions on Interactive Intelligent Systems*, 6(4), November 2016. doi:10.1145/2963106.

74 Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4–5):441–504, 2012. doi:10.1007/s11257-011-9118-4.

75 Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18:140–181, 2009. doi:10.1007/s10618-008-0114-1.

76 Carol C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991. doi:10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-\%23.

77 Monica Landoni, Mohammad Aliannejadi, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. Right way, right time: Towards a better comprehension of young students' needs when looking for relevant search results. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 256–261, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3450613.3456843.

78 Monica Landoni, Davide Matteri, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. Sonny, cerca! evaluating the impact of using a vocal assistant to search at school. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 101–113. Springer, 2019. doi:10.1007/978-3-030-28577-7_6.

79 Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/800de15c79c8d840f4e78d3af937d4d4-Paper.pdf.

80 Jinggui Liang, Yuxia Wu, Yuan Fang, Hao Fei, and Lizi Liao. A Survey of Ontology Expansion for Conversational Understanding. In H. Bouamor, J. Pino, and K. Bali, editors, *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 18111—-18127. Association for Computational Linguistics, USA, 2023. doi:10.18653/v1/2024.emnlp-main.1006.

81 Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. LLM-REDIAL: A large-scale dataset for conversational recommender systems created from user behaviors with LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 8926–8939, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-acl.529.

82 Aldo Lipani, Ben Carterette, and Emine Yilmaz. How am I doing?: Evaluating conversational search systems offline. *ACM Transactions on Information Systems*, 39(4), August 2021. doi:10.1145/3451160.

83 Zeyang Liu, Ke Zhou, and Max L Wilson. Meta-evaluation of conversational search evaluation metrics. *ACM Transactions on Information Systems*, 39(4), September 2021. doi:10.1145/3445029.

84 William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988. doi:10.1515/text.1.1988.8.3.243.

85 Gary Marchionini. Exploratory Search: From Finding to Understanding. *Commun. ACM*, 49(4):41–46, 2006. doi:10.1145/1121949.1121979.

86 D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, 13(3):334–359, 2002. doi:10.1287/isre.13.3.334.81.

87 Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: a human generated MAchine Reading COmprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne,

editors, *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

88 Elaine Ostroff. Universal design: an evolving paradigm. In *Universal Design Handbook*, chapter 1. McGraw-Hill, New York, NY, USA, 2011.

89 Paul Owoicho, Jeff Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R. Trippas, and Svitlana Vakulenko. TREC CAsT 2022: Going beyond user ask and system retrieve with initiative and response generation. In Ian Soboroff and Angela Ellis, editors, *Proceedings of the Thirty-First Text REtrieval Conference, (TREC 2022)*, volume 500-338 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2022. URL: https://trec.nist.gov/pubs/trec31/papers/Overview_cast.pdf.

90 Patrick Cheong-Iao Pang, Shanton Chang, Karin Verspoor, and Jon Pearce. Designing health websites based on users' web-based information-seeking behaviors: A mixed-method observational study. *Journal Medical Internet Research*, 18(6):e145, June 2016. doi:10.2196/jmir.5661.

91 Patrick Cheong-Iao Pang, Karin Verspoor, Shanton Chang, and Jon Pearce. Conceptualising health information seeking behaviours and exploratory search: result of a qualitative study. *Health and Technology*, 5:45–55, 2015.

92 Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. Mobile-Based Experience Sampling for Behaviour Research. In Marko Tkalčič, Berardina De Carolis, Marco de Gemmis, Ante Odić, and Andrej Košir, editors, *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, pages 141–161. Springer International Publishing, Cham, 2016. doi:10.1007/978-3-319-31413-6_8.

93 G Penha and C Hauff. Challenges in the evaluation of conversational search systems. In *KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption, KDD-Converse 2020*, volume 2666, 2020.

94 Gustavo Penha, Alexandru Balan, and Claudia Hauff. Introducing MANtIS: a novel multi-domain information seeking dialogues dataset. *CoRR*, abs/1912.04639, 2019. arXiv:1912.04639.

95 Maria Soledad Pera, Theo Huibers, Emiliana Murgia, and Monica Landoni. Some things never change: Overcoming persistent challenges in children IR. In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne, editors, *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 3107–3109. ACM, 2025. doi:10.1145/3726302.3730270.

96 Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3020165.3020183.

97 Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. A Survey on Asking Clarification Questions Datasets in Conversational Systems. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 2698–2716. Association for Computational Linguistics, USA, 2023. doi:10.18653/v1/2020.acl-main.363.

98 Robert Ramberg and Klas Karlgren. Fostering superficial learning. *Journal of Computer Assisted Learning*, 14(2):120–129, 1998. doi:10.1046/j.1365-2729.1998.1420120.x.

99 Tetsuya Sakai. SWAN: A generic framework for auditing textual conversational systems. *CoRR*, abs/2305.08290, 2023. doi:10.48550/arXiv.2305.08290.

100 Tetsuya Sakai, Douglas W. Oard, and Noriko Kando, editors. *Evaluating Information Retrieval and Access Tasks – NTCIR's Legacy of Research Impact*, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2021.

101 Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends™ in Information Retrieval*, 4(4):247–375, 2010. doi:10.1561/1500000009.

102 Michael Schulte-Mecklenbeck, Joseph G. Johnson, Ulf Böckenholt, Daniel G. Goldstein, J. Edward Russo, Nicolette J. Sullivan, and Martijn C. Willemsen. Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, 26(5):442–450, 2017. doi:10.1177/0963721417708229.

103 John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969. doi:10.1017/CBO9781139173438.

104 Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. Reliable LLM-based user simulator for task-oriented dialogue systems. In Yvette Graham, Qun Liu, Gerasimos Lampouras, Ignacio Iacobacci, Sinead Madden, Haider Khalid, and Rameez Qureshi, editors, *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat*, SCI-CHAT 2024, pages 19–35, St. Julians, Malta, March 2024. Association for Computational Linguistics. URL: https://aclanthology.org/2024.scichat-1.3.

105 C. Shah and E. M. Bender. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? *ACM Transactions on the Web (TWEB)*, 18(3):33:1–33:24, August 2024. doi:10.1145/3649468.

106 Jack Sidnell and Tanya Stivers. *The Handbook of Conversation Analysis*. Blackwell, 2012. doi:10.1002/9781118325001.

107 Stefan Sitter and Adelheit Stein. Modeling the illocutionary aspects of information-seeking dialogues. *Information Processing & Management*, 28(2):165–180, 1992. doi:10.1016/0306-4573(92)90044-Z.

108 Neil R Smalheiser. Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology*, 63(2):218–224, 2012. doi:10.1002/asi.21599.

109 Neil R. Smalheiser. Rediscovering don swanson:the past, present and future of literature-based discovery. *Journal of Data and Information Science*, 2(4):43–64, 2017. doi:10.1515/jdis-2017-0019.

110 Damiano Spina, Johanne R. Trippas, Paul Thomas, Hideo Joho, Katriina Byström, Leigh Clark, Nick Craswell, Mary Czerwinski, David Elsweiler, Alexander Frummet, Souvick Ghosh, Johannes Kiesel, Irene Lopatovska, Daniel McDuff, Selina Meyer, Ahmed Mourad, Paul Owoicho, Sachin Pathiyan Cherumanal, Daniel Russell, and Laurianne Sitbon. Report on the future conversations workshop at CHIIR 2021. *SIGIR Forum*, 55(1), July 2021. doi:10.1145/3476415.3476421.

111 Amanda Spink, Minsoo Park, Bernard J. Jansen, and Jan Pedersen. Multitasking during Web search sessions. *Information Processing & Management*, 42(1):264–275, 2006. doi:10.1016/j.ipm.2004.10.004.

112 Manfred Stede and Jodi Schneider. *Argumentation Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2018. doi:10.2200/S00883ED1V01Y201811HLT040.

113 Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics, 2019. doi:10.18653/V1/P19-1355.

114 Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1003.

115 Özge Sürer, Robin Burke, and Edward C Malthouse. Multistakeholder recommendation with provider constraints. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 54–62, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3240323.3240350.

116 Don R Swanson. Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*, 38(4):228–233, 1987. doi:10.1002/(SICI)1097-4571(198707)38:4<228::AID-ASI2>3.0.CO;2-G.

117 Robert S. Taylor. The process of asking questions. *American Documentation*, 13(4):391–396, 1962. doi:10.1002/asi.5090130405.

118 Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In Elizabeth Dykstra-Erickson and Manfred Tscheligi, editors, *Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI 2004)*, pages 415–422. Association for Computing Machinery, 2004. doi:10.1145/985692.985745.

119 Helma Torkamaan, Mohammad Tahaei, Stefan Buijsman, Ziang Xiao, Daricia Wilkinson, and Bart P. Knijnenburg. The Role of Human-Centered AI in User Modeling, Adaptation, and Personalization—Models, Frameworks, and Paradigms. In Bruce Ferwerda, Mark Graus, Panagiotis Germanakos, and Marko Tkalčič, editors, *A Human-Centered Perspective of Intelligent Personalized Environments and Systems*, pages 43–84. Springer Nature Switzerland, Cham, 2024. doi:10.1007/978-3-031-55109-3_2.

120 Johanne R Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. What do users really ask large language models? an initial log analysis of Google Bard interactions in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3626772.3657914.

121 Johanne R Trippas, Luke Gallagher, and Joel Mackenzie. Re-evaluating the command-and-control paradigm in conversational search interactions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3627673.3679588.

122 Johanne R. Trippas, Damiano Spina, and Falk Scholer. Adapting generative information retrieval systems to users, tasks, and scenarios. In Ryen W. White and Chirag Shah, editors, *Information Access in the Era of Generative AI*, pages 73–109. Springer Nature Switzerland, Cham, Switzerland, 2025. doi:10.1007/978-3-031-73147-1_4.

123 Johanne R Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 57(2), 2020. doi:10.1016/J.IPM.2019.102162.

124 Robin Ungruh and Maria Soledad Pera. Ah, that's the great puzzle: On the quest of a holistic understanding of the harms of recommender systems on children. *CoRR*, abs/2405.02050, 2024. doi:10.48550/arXiv.2405.02050.

125 Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. QRFA: A data-driven model of information-seeking dialogues. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval - 41st European Conference on IR Research (ECIR 2019)*, volume 11437 of *Lecture Notes in Computer Science*, pages 541–557. Springer, 2019. doi:10.1007/978-3-030-15712-8_35.

126 R. W. White and C. Shah, editors. *Information Access in the Era of Generative AI*, volume 51 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2025.

127 Ryen W. White. Tasks, copilots, and the future of search: A keynote at sigir 2023. *SIGIR Forum*, 57(2), January 2024. doi:10.1145/3642979.3642985.

128 Ryen W White and Resa A Roth. *Exploratory search: Beyond the query-response paradigm*. Springer, Cham, Switzerland, 2009. doi:10.1007/978-3-031-02260-9.

129 John M Wiemann. Explication and test of a model of communicative competence. *Human communication research*, 3(3):195–213, 1977. doi:10.1111/j.1468-2958.1977.tb00518.x.

130 Martijn C Willemsen and Eric J Johnson. Visiting the decision factory: observing cognition with MouselabWEB and other information acquisition methods. In Michael Schulte-Mecklenbeck, Anton Kuehberger, and JohnsonJoseph G., editors, *A handbook of process tracing methods for decision research: a critical review and user's guide*, pages 21–42. Psychology Press, 2011. URL: http://repository.tue.nl/699163.

131 Robert L Winkler and Allan H Murphy. Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance*, 10(2):252–270, 1973. doi:10.1016/0030-5073(73)90017-2.

132 A Wayne Wymore. *Systems Engineering Methodology for Interdisciplinary Teams*. Wiley, New York, NY, USA, 1976.

133 Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3–4):244–456, August 2023. doi:10.1561/1500000081.

134 Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Computing Surveys*, 55(8), 2022. doi:10.1145/3556536.

135 Savvas Zannettou, Olivia Nemes Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P. Gummadi, Elissa M. Redmiles, and Franziska Roesner. Analyzing user engagement with tiktok's short format video recommendations using data donations. In Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas, Max L. Wilson, Phoebe O. Toups Dugas, and Irina Shklovski, editors, *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI 2024, pages 731:1–731:16. Association for Computing Machinery, 2024. doi:10.1145/3613904.3642433.

136 Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. The SimIIR 2.0 framework: User types, markov model-based interaction simulation, and advanced query generation. In Mohammad Al Hasan and Li Xiong, editors, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM 2022, pages 4661–4666. Association for Computing Machinery, 2022. doi:10.1145/3511808.3557711.

137 Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. USimAgent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 2687–2692, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3626772.3657963.