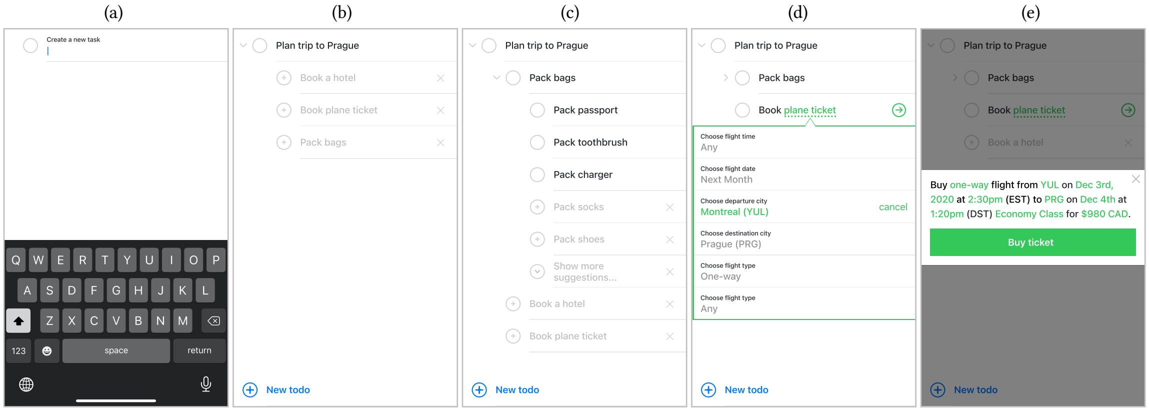
ToBeDone: Exploration of a todo application paradigm for general purpose intelligent assistance

ANONYMOUS AUTHOR(S)



(a) Create a new task

(b) Plan trip to Prague
Book a hotel
Book plane ticket
Pack bags
New todo

(c) Plan trip to Prague
Pack bags
Pack passport
Pack toothbrush
Pack charger
Pack socks
Pack shoes
Show more suggestions...
Book a hotel
Book plane ticket
New todo

(d) Plan trip to Prague
Pack bags
Book plane ticket
Choose flight time: Any
Choose flight date: Next Month
Choose departure city: Montreal (YUL) cancel
Choose destination city: Prague (PRG)
Choose flight type: One-way
Choose flight type: Any
New todo

(e) Plan trip to Prague
Pack bags
Book plane ticket
Book a hotel
Buy one-way flight from YUL on Dec 3rd, 2020 at 2:30pm (EST) to PRG on Dec 4th at 1:20pm (DST) Economy Class for $980 CAD.
Buy ticket
New todo

Fig. 1. ToBeDone interaction design for the "plan a trip" scenario: (a) user creates a new task called "Plan trip to Prague"; (b) the AI system adds suggestions as subtasks; (c) clicking a suggestion like "Pack bags" adds a subtask; (d) green annotations indicate the level of AI uncertainty, the user review and modify task parameters inferred by the AI system; (e) the user then asks the AI to perform the task, in this case to buy a plane ticket.

We explore the concept of an AI-augmented todo application as a non-anthropomorphic Intelligent Assistance paradigm to provide users with transparent, controllable, and integrated help for every day tasks. We focus on the interface and interaction design for such an application. A first study evaluates and refines two initial prototype variations. A second study compares the refined todo prototype with the established paradigm of an anthropomorphic voice-based Virtual Assistant. Results show that while the virtual assistant is preferred for the simplest scenario we tested, our todo paradigm was more transparent, controllable, and preferred for the more complex scenarios. Our work introduces an alternative interaction paradigm for general purpose AI that prioritizes predictability, control, and augmentation.

Additional Key Words and Phrases: artificial intelligence, interaction, todo

1 INTRODUCTION

The virtual assistant (VA) paradigm popularized by systems like Siri, Google Assistant, Alexa, and Cortana, represents artificial intelligence (AI) processes as an anthropomorphic entity living inside a personal device. The goal is to provide people with general-purpose assistance for everyday tasks. A user can converse with their VA to do common things like set an alarm, call a friend, or check the weather. They can even make some specialized requests like find out what plane is currently flying over their head, provided they are aware of this capability.

The open-ended VA paradigm is flexible and general since it is not limited to predefined actions accessible only through specific menus and widgets in an interface. Instead, users can request anything they can formulate through natural language. The downside is that as advanced as the underlying AI system is, many requests will be far beyond its intrinsically limited capabilities. The human-like behaviour of VAs is thought to make interaction more natural [5] and to improve trust in autonomous processes [10]. This argument is often grounded in Nass et al.'s work arguing computers are considered social actors [35]. However more recently, this has been questioned by Shechtman and Horowitz [46], who exposed clear behaviour differences when people interact with other humans, compared to computers. In addition, Luger and Sellen [30] observed user expectations are often greatly misaligned with the actual system capabilities, leading to frustration. As a result, users relegate VA usage to limited occasions, when alternatives are not possible, such as when their hands are occupied, or for simple use cases, like setting timers [29, 30]. Arguably, the diversity and frequency of intelligent assistance accessed through VAs falls well below the benefits AI could offer.

We reimagine interaction with a general purpose virtual assistant by departing from an anthropomorphised conversational dialog, and instead explore intelligent assistance that is tightly integrated into existing general purpose applications. In this paper, we investigate how AI can augment a todo list application, a tool people already use to record their goals. We call this interaction paradigm *ToBeDone*. Based on guidelines for Human-AI Interaction and a review of relevant literature, we devised three design values for our approach: predictability of AI behaviour through transparency, user control of the result, and augmentation by integrating AI into existing user workflows.

Following our design values, we designed two initial mockups of AI-augmented todo list applications. These mockups are evaluated, and fine-tuned, during a first user study testing three different scenarios: placing a phone call, preparing a meal, and planning a trip. Following this first study, we further refined our design mockup, compare it to the traditional VA paradigm in a second user study using the same three scenarios. To make interaction paradigms comparable, the simulated AI exhibited the exact same behaviour in both conditions. Results show that while the VA paradigm is preferred for the simplest scenario, our ToBeDone paradigm was more transparent, controllable, and preferred for the more complex ones. We discuss AI interaction and reflect on our design values in light of these results.

We make two contributions:

- A novel interaction paradigm to create AI-augmented systems that focus on three values: transparency, control, and augmentation,
- The results of two qualitative studies investigating the usability of our approach, and comparing it to traditional VAs.

2 BACKGROUND AND RELATED WORK

Interaction with AI systems can be challenging for users. They can be frustrated by inaccuracies even when rare, confused by capabilities and inabilities [29], feel distrustful [30], and be unwilling to relinquish control [45]. These issues can even push users to disregard AI features altogether [15, 45]. Various approaches to AI interaction have been used to alleviate these issues.

2.1 Anthropomorphic Virtual Assistants

AI systems often implement Beaudouin-Lafon's computer-as-partner interaction principle [4]: the intelligent behaviour materializes as an anthropomorphic entity acting as a virtual personal assistant with a human-like personality. Apple Siri, Google Assistant, Amazon Alexa, and Microsoft Cortana all testify to industry enthusiasm for the VA paradigm. However, the approach largely predates modern assistants: one may recall "Clippy" for example: a VA personified as a living paper clip was part of Microsoft Word from 1997 to 2003.

The goal of many previous works is to make a VA behave like a human to increase user engagement, reduce frustration [8, 21, 25], and improve trust [6, 10, 19, 39]. Since interacting with other humans is often easier than interacting with computers, anthropomorphization is thought to improve the communication process [5]. For example, Maes and Kozierok [34] use simple human caricatures to convey system states like "thinking", "suggesting", or "unsure".

Bickmore and Picard [6] argue that by establishing a long-term relationship with the system, both entities will eventually become familiar and trust each other. Or again, Cassell and Bickmore [10] aim at making the artificial agent behave in a way that promotes trustability through small talk. Anthropomorphism can manifest itself in various degrees. In its simplest form, chatbots like Eliza [57] communicate with users with natural language as text. At the opposite end of the spectrum, embodied assistants like Rea [11] can make use of voice intonation, gestures, gaze, and postures [9, 13]. Cassell [9] argues intelligence needs to be embodied because users need to physically locate it to trust it.

Yet the use of anthropomorphism in user interfaces is contested. For example, Walker et al. [56] argue that humanizing an interface adds superfluous information that needs to be processed by users, which distracts them from their primary task. Shneiderman [51, 53] argues it reduces the feeling of being in control, and suggest that presenting computers as partners is misleading as humans remain solely responsible for the technology they use and its results [52]. Other works highlight instances where social expressions were considered “burdensome” [29]. Cramer et al. [12] even found it made the system appear less trustworthy: their participants were dubious of empathizing utterances because they did not believe the system could actually empathize. Höök [22] argues that anthropomorphism tends to mislead users about the system capabilities, increasing Norman’s gulf of evaluation [36], and leading to frustration when expectations are not met. Luger and Sellen [31] reached similar conclusions with commercial VAs, noting that such deception can be conveyed by minor and subtle elements of the interaction, such as humour.

2.2 AI-augmented systems

Of course, AI systems do not need to communicate with users using natural language, or to be associated with an anthropomorphic figure to be useful. These systems generally implement Beaudouin-Lafon’s computer-as-tool principle [4]. The AI augments the graphical user interface. A familiar example are the word suggestions that appear on a phone keyboard to provide typing shortcuts [44]. Google Smart Compose[1] shows how even more AI automation can be integrated into an interface by proposing entire sentences at the insertion point of an email composition window. Augmentation enables in-situ AI assistance, for example SmartEye [32] learns from user’s preference and aid photo composition at capture time. On the other hand, loosely integrated AI systems with their own separate interface tend to be used less. Indeed, Yang et al. [58] noticed decision support tools are rarely used despite compelling evidence of their effectiveness. Seeking tighter integration within the existing practitioner workflow, they incorporated AI predictions into slides used for decision meetings. Finally, augmentation may contribute to empower user instead of taking control away from them. For example, Rescribe [40] uses AI to assist users in generating and refining video descriptions. User may iterate on the system’s result, and vice-versa, enabling great controllability.

However, AI augmentations generally focuses on one specific task, such as improve photo composition. User interaction is often minimal, such as pressing the tabulation key to accept the proposed sentence completion. In contrast, virtual assistants like Apple Siri or Microsoft Cortana are designed to be general purpose: they may be used for a large variety of tasks such as triggering a call, checking the weather, or setting up a reminder. We are interested in the design of integrated non-anthropomorphic interaction with general purpose AI: an intelligent personal assist*ance*, but not an intelligent personal assist*ant*. We investigate if moving away from anthropomorphism for general purpose intelligent personal assistance may help alleviate associated issues, and provide useful access to AI systems on which users rely.

[1] https://ai.googleblog.com/2018/05/smart-compose-using-neural-networks-to.html

2.3 Human-AI interaction

AI systems have the potential to reduce the collective workload for humanity. Automation communities generally aim at creating *partners* able to carry out tasks on a user's behalf [33]. In contrast, HCI communities emphasize the design of *tools* to augment human intellect [49]. This subtle difference sometimes leads to misunderstandings between the two communities [53], in particular about *AI autonomy* and *user controllability*. AI research often strives to make systems more autonomous, effectively reducing what is left for users to do by themselves. In exchange, the system often requires users to release some control on the outcome of the task. For example, it can be tedious to correct which contact to call when a VA misunderstands a name. This goes against HCI principles arguing that users have a strong need to be in control of their task [49, 50]. In fact, results indicate users are more satisfied with a highly controllable system than a highly accurate one [45], and that a high level of system autonomy may actually increase user workload because they have to monitor the behaviour of an unpredictable system [54].

Automation autonomy does not necessarily imply less controllability, and it is possible, albeit difficult, to design systems that integrate both high levels of system autonomy and control for users [52]. Shneiderman [52] uses the example of Patient Controlled Analgesia (PCA) devices. Sophisticated PCA models enable users to get more medicine by pressing a button, but rely on an AI system to deliver a quantity appropriate for the patient, enough to relieve pain while also preventing overdoses. Related to this are Horvitz's formative work on mixed-initiative [23] that investigates methods to make VAs more controllable by adding direct manipulations to the conversation [51]. We aim at exploring new ways to interact with intelligent assistance that would enable both high levels of AI autonomy and user control.

Trust is another central concept of Human-AI interaction, and is the focus of many work as surveyed by Hoff and Bashir [20]. The concept is borrowed from social science, and definitions often differ. Lee and See [28] define it as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". It is arguably impacted by the accuracy of the automation, but is hard to predict as there are many other factors. For example, Efendić et al. [16] showed that users tend to trust automation more the faster it presents predictions. As discussed earlier, anthropomorphism is often used to improve trust [6, 10], although this can be counterproductive [12, 29]. Trust influences *Reliance* but does not determine it [15, 28]: other factors like controllability may also impact how much users rely on AI [45]. Perhaps surprisingly Greater trust is not necessarily a sign of more effective design [28]. Indeed, *overtrust* may encourage people to rely on AI inappropriately. Consider the example of drivers spending more time on non driving-related tasks in semi-autonomous vehicles [27]. Lee and See [28] argue trust must be *calibrated* for appropriate reliance.

Our work builds heavily on guidelines for AI interaction [2, 22–24, 37]. In particular, those Amershi et al. [2] review as fundamental recommendations, including showing contextually relevant information, being transparent about what the system can do, and how well it can do it. These two last points are related to AI *transparency* and *explainability* that have been the focus of many previous works [1, 26, 42, 43]. Explaining AI behaviour remains tremendously challenging. In this work, we focus on making the result of AI processes transparent and predictable to the user.

3 DESIGN VALUES

We devised three design values: predictability, control, and augmentation. These values address obstacles to useful intelligent assistance in everyday tasks identified in our literature review, and are informed by AI interaction guidelines.

3.1 Predictability

AI behaviour should be transparent, meaning results should at the least be predictable even if not always understandable. Feedforward mechanisms [14, 55] show users what the result of an action will be and are an important characteristic that is lacking in the VA paradigm. As an example, the AI system could pre-calculate a result and present it for the user to accept or not, instead of requiring the user to fully execute the AI system to judge result quality and AI capability. A simple form of this approach is used for word suggestions [3] where predictions are shown either as buttons on top of the keyboard, or after the insertion cursor, making behaviour transparent and the result entirely predictable. Predictability may help alleviate issues related to trust since one does not need to trust an AI system if results are visible before committing to their execution.

Transparency may incur a cognitive cost during normal use when users are forced to consider whether to use the automation or not. It can make the AI counterproductive and burden users instead of serving them. As an example, previous work observed that word suggestions may actually slow down users [41, 44]. Consequently, appropriate feedforward must be designed to balance AI services and cognitive cost.

3.2 Control

When AI accuracy is high and the system is almost always able to produce a satisfactory result, it may be tempting to neglect a user interface for users to refine and correct results. However, the inability to fix results may lead to user frustration when inaccuracy unavoidably occurs [45]. For example, image stitching algorithms are generally able to properly assemble seamless panoramas, but they occasionally fail and render distorted figures and other undesirable artifacts. In many cases, the interface provides no way for users to salvage their work, an example of poor controllability.

While AI autonomy and user control are often presented as two extremes of a single continuous dimension [47, 48], they are actually orthogonal [52], and it is possible to design a system exposing high levels of both. While the AI system should attempt as much as possible to guess user intent, the user interface should always provide users with the opportunity to refine their need and adjust incorrect predictions. As an example, it could be possible to fix a panorama by dragging over stitching errors [18].

3.3 Augmentation

VAs tend to have their own dedicated user interface, disconnected from the user's work and routines. Unfortunately, users often do not recognize they could benefit from AI help [58]. In addition, firing up a separate application, like most VA implementations, implies a change of context incurring a cognitive load. We aim to explore a different path where the intelligent assistance is an optional augmentation of existing application functionality. The AI integrates with existing direct manipulation mechanisms [49], predominant in graphical user interfaces. It augments them instead of replacing them. Previous work indicates a strong tendency for users to rely on manual controls, even when the AI is highly accurate [45]. We assume the manual path will remain preferred and most often used, so users should never be forced to rely on the AI to complete their task. The AI system should not impact the usability of manual controls, but strive to accompany and improve them instead. This enables users to leverage AI when it is accurate, for example by letting AI book a train ticket for a recurring trip, but also to easily ignore it and complete their task as usual if they wish.

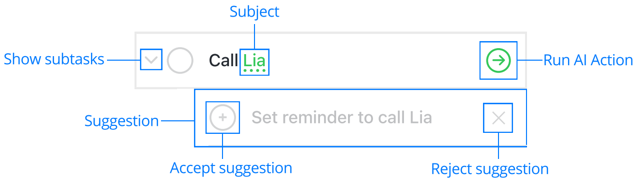


Fig. 2. Tasks and suggestions in ToBeDone with the Hypertext Annotations Variation.

4 TOBEDONE INTERACTION DESIGN

To investigate non-anthropomorphic intelligent assistance that is tightly integrated into existing general purpose applications, we augment a smartphone todo application we call ToBeDone. A todo application exposes simple premises: users create, edit, organize, complete, and delete tasks in the form of todo items. Todo items can also be nested inside others to form subtasks. A todo application is an interesting basis for AI augmentation that helps users achieve everyday tasks. Appropriately, it is often the hub of an organized user's life were intents and tasks are recorded, reflected upon, initiated, and eventually completed.

We augmented two todo functions: task creation and task completion. For task, the AI suggests relevant tasks the user can create, which we call suggestions. For task completion, the AI suggests relevant actions it can take to complete a task, which we called actions. We describe the interaction design to realise these augmentations below.

There is much information in a todo application that the underlying AI system could leverage, including user tasks and which suggestions are accepted or rejected. However, we focus on interaction design, and leave the functional AI system as a complementary, but very different topic.

4.1 Suggestions

Through suggestions, the underlying AI system can help the user complete their goals by suggesting contextually relevant tasks. We augment the familiar interaction of task creation to let the user easily invoke and dismiss suggestions.

Task suggestions appear as greyed-out tasks that user can accept or reject (see Figure 2). Once the user accepts a suggestion, it behaves like a user-created task. To the right end of the suggestion is a close icon which the user can press to remove it. Suggestions may appear as tasks or subtasks. For example, if the user adds a task "Create packing list for trip", the AI would suggest possible items to pack in the form of subtask suggestions such as "pack socks" or "pack umbrella". The user can always choose to ignore suggestions and use the system like a normal todo application.

4.2 Actions

An action is one of a series of steps the AI can execute in order to assist the user in completing a task. When the user creates a task, the underlying AI system may identify a possible action it could perform, similar to how a virtual assistant identifies an action from a spoken command. As an example, the AI would interpret the task "Call Lia Kane" as an intent of calling a contact and suggest the action of calling the phone number associated with "Lia Kane".

Figure 3 illustrates ToBeDone action icon states. The presence of a green arrow icon to the right end of a task indicates it is actionable. A solid green icon indicates the AI has all information it needs to run the action without making assumptions about the user's intent (Figure 3a). An outlined green icon indicates the AI can execute the action, but it made assumptions that have not been confirmed by the user (3b). An outlined orange question mark icon indicates the AI requires more information from the user to execute an action (3c). Lastly, an outlined grey icon indicates the AI cannot execute the action to complete the task (3d).

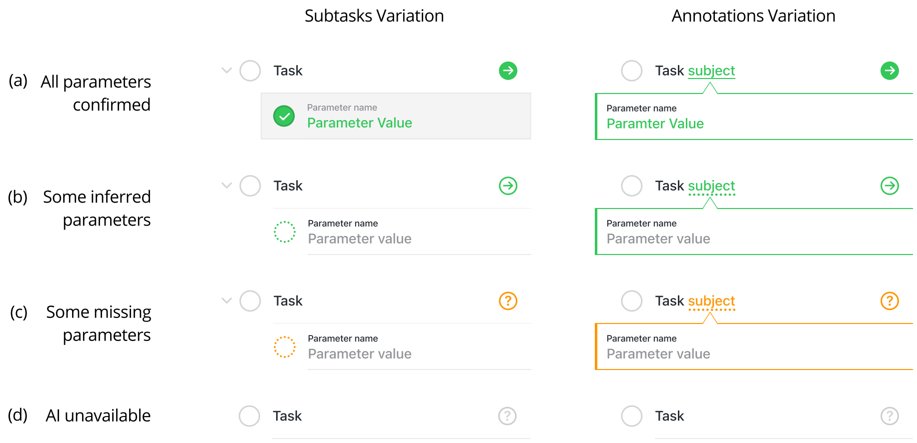


Fig. 3. Action and action parameters with ToBeDone.

When the user taps the green action icon of a task, a panel appears explaining what the AI is about to execute and asks for the user's confirmation. The action is composed of various AI-inferred parameters. Once the user confirms the action, the AI executes it and marks the task as complete. When a user marks a task as complete it is indicated with a solid blue check mark. When the AI executes an action for a task, it marks it as complete using a solid green check mark to notify the user of changes its made. The user can revisit the action panel to see what actions the AI executed. Like suggestions, actions can be easily ignored and the user can use the application as a normal todo application.

4.3 Action Parameters

Action parameters provide *feedforward* on the AI action, making its result *predictable*. Because user can refine them when AI predictions are inaccurate, it also gives them *control*.

Consider the simple but ambiguous task "Book train ticket". Because predictions are transparent and easy to refine by users, ToBeDone encourages the underlying AI system to always guess as much as possible, even if confidence is low. In this case, it could pick a destination, and a time based on user's calendar, and decide on a price range according to the user's habits. When AI inferences happen to be all correct, this streamlines the user's workflow enabling them to complete the task with a tap on the green action icon. However, the user can override any parameter with their own value, triggering an update of the corresponding action.

4.4 Design Variations

The main challenge in augmenting an interface with AI is to subtly display additional information while minimizing disruptions. A balance must be struck between the cost of disruptions, and the benefits provided by the underlying AI system. We constructed two approaches to address this challenge which resulted in two different versions of our design. They differ in how task parameters are presented to the user: subtasks or annotations.

4.4.1 Subtasks. This design shown in Figure 3 left leverages the interaction principles already existing within a todo list application. It presents action parameters as modifiable AI-generated subtasks. For example, if a user creates a todo to book a flight, the AI may insert "Choose departure date", "Choose destination", and "Choose maximum acceptable price". These integrate well with the todo paradigm because they are effectively intermediate steps user needs to consider anyway before the overarching task of booking a flight can be accomplished, and are formulated as such. Modifying these generated subtasks update the corresponding parameter, and allows refining the action.

Similar to a task action, the checkbox associated with a parameter subtask indicates its state. A dotted orange circle means the AI cannot infer a value and is not able to provide assistance without further information. A dotted green circle means the AI inferred a value, but it has not been confirmed by the user. A solid green circle indicates the user has set the parameter, or has confirmed an inferred value.

4.4.2 Hypertext Annotation. This version shown in Figure 3 right adds annotations to a user's task, and provides form-like widgets to visualize and edit action parameters. Similar to spelling errors in word processing software, the underlying AI system can underline text that are descriptors related to the action parameters. Selecting an annotation displays the action parameters, and they can be modified in the same way as the subtasks version. As an example, if the user creates a todo with the text "Book a flight", the AI may recognize "flight" as a relevant descriptor and underline it. Selecting this annotation opens a widget with the same information in the subtasks version that can be similarly filled or modified.

The underline is rendered in different way to convey AI confidence. The same visual language is used as the circles of the subtask version, for example a solid dotted green circle corresponds to a solid dotted green underline.

5 STUDY 1: USABILITY EVALUATION

The goal of this study is to test and refine the interaction designs of our ToBeDone intelligent assistance paradigm for later comparison with the VA paradigm. We test interactive prototypes for both ToBeDone design variations with three different scenarios: phone call, meal planning, and trip planning. Using a think-aloud protocol, we ask participants to interact with the prototypes on their own smartphone. We supply a goal and series of prompts which they follow to complete each scenario. After, we ask a series of questions pertaining to usability and the ToBeDone paradigm. The study aims at identifying pitfalls and design successes before converging on a final design.

5.1 Participants

We analyzed the data of 8 participants: ages 19 to 26 (M = 21.3, SD = 2.1); 6 self-reported as male and 2 as female. 3 had written AI code or followed a course about AI, 4 had read about AI, and 1 knew little about it. 5 participants had used a VA at least once a week during the past 30 days. Remuneration was a $10 gift card for a 1 hour study.

5.2 Procedure

The study was conducted online, with the participant using interactive prototypes created in Figma[2], loaded as a web page on their phone to simulate an application experience. The facilitator communicated using video conferencing, and they were able to monitor participant actions and progress in the Figma editor. Audio was recorded during the study for transcription and analysis.

Each session started with informed consent, and a short demographic questionnaire. The facilitator then read an introduction script outlining the study purpose and task. However, they did not answer questions or provided explanations about the way mockups are supposed to be operated. Participants then completed each of the three scenarios using both prototypes. They followed a series of spoken prompts provided by the facilitator, for example "Call your friend Lia Kane" or "Book a plane ticket leaving from Montreal on Dec 3rd 2020". Participant were instructed

[2] https://www.figma.com

to explain their thought process as they completed each prompt. After the participant had completed a scenario with both design variations, the facilitator gathered feedback in a short structured interview ("How do you feel about the two different versions of the app?"), inquired about usability ("Was there anything confusing, difficult, or unclear to you when using the app?"), and preferred prototype ("Which version of the app do you prefer and why?").

After all three scenarios were completed with both design variations, the facilitator conducted a short structured interview in which the participant was asked to make a subjective comparison to a regular todo list application ("How would you compare this experience to a regular todo app?"), and a subjective comparison to current VAs ("How do you feel about completing these tasks using the app versus using a virtual assistant?").

We followed a Concurrent Think Aloud (CTA) protocol as a way to detect usability problems. Participants were asked to verbalize all thoughts as they interacted. Following Olmsted-Hawala et al.'s study [38], we implemented Boren and Ramey's speech communication-based CTA protocol [7]. Boren and Ramey's approach varies from Ericsson and Simon's original CTA methodology [17] in that the facilitator follows the participant's utterances with regular acknowledgement or response tokens instead of refraining from any communication. These include a usual "keep talking" if the participant fell silent, but also verbal feedback in form of "um-um", "oh", "okay" to keep the participant talking. The facilitator may also include a questioning tone picking up on the last word uttered by participants after 15 seconds of silence: for example if a participant says "this is weird…," the facilitator may say after a pause, "Weird?".

5.3 Prototypes and Scenarios

We used Figma to explore, iterate on, and build our prototypes, and run the study. Interaction design and usability was our main focus, so the prototypes only included the intended path for each scenario and design variation. We created three different scenarios to use as study tasks: CALL, MEAL, and TRIP.

The scenario CALL was intended to be the simplest (Figure 6). Participants first set a reminder to call a friend, upon which the AI would provide a shortcut to make the call, but it makes an incorrect assumption about the recipient. Participants had to correct this assumption before making the call. This scenario investigated using ToBeDone in the context of a simple task, participant's ability to perceive AI inaccuracies, and to fix them.

The second scenario was MEAL (see Figure 4). The participant decides to cook at home rather than going to a restaurant, they refined, navigated, and selected meal suggestions, check a generated ingredient shopping list, and consult a recipe. This scenario made participants move through a series of heterogeneous sub-tasks and different type of functionalities AI could enable: user intent refinement into actionable items, suggestions, shopping list generation, and document consultation. It also included prediction inaccuracies investigating how participants may refine the information provided to the system.

The last and most complex scenario was TRIP (see Figure 1). The participant manually adds items to pack, uses automatic suggestions for items to pack, refines plane ticket details, and books a plane ticket. It investigated how participants can use the system using both manual and automatic functionalities, and achieve tasks with many variables.

We provide videos demonstrating the study scenarios as supplementary materials.

5.4 Design

This is a within subjects design with two prototype conditions for our interaction design variations: SUBTASKS is the design using subtasks as action parameters, and ANNOTATE is the design using hypertext annotations. Each was evaluated in the context of 3 scenarios: CALL, MEAL, TRIP. The order of design variation was counterbalanced, scenario order was fixed according to the expected increase in complexity: CALL, MEAL, then TRIP.

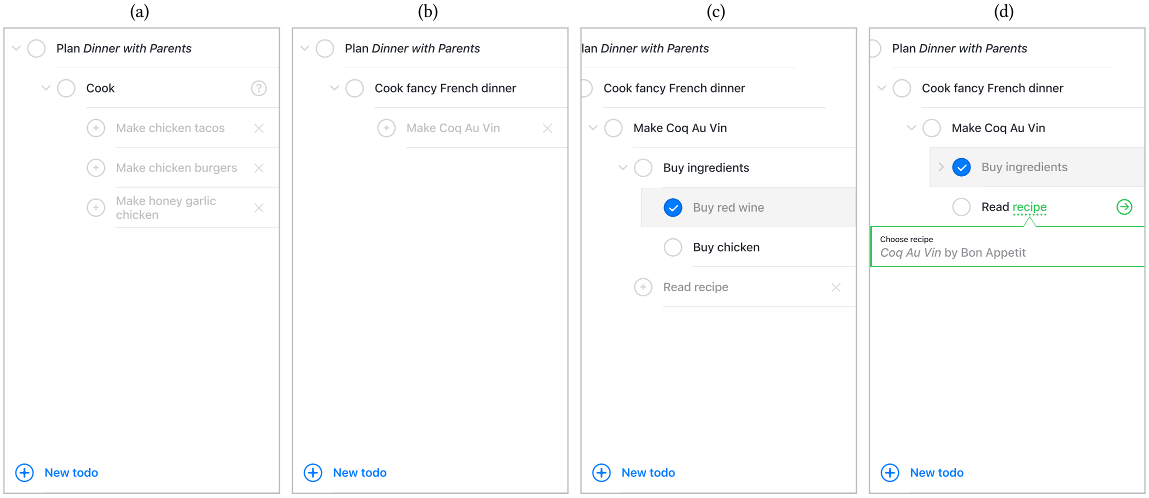


Fig. 4. Planning a meal with ToBeDone's subtasks variation. (a) Check AI meal suggestions; (b) Refine intent; (c) Buying ingredients following AI suggestions, the interface slides left to focus on subtasks; (d) Check recipe following AI suggestion.

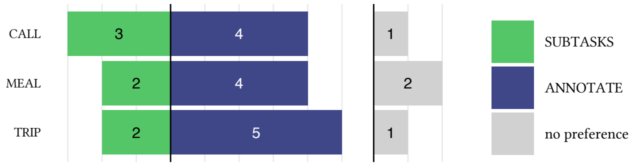


Fig. 5. Number of participants preferring each design variation per scenario.

5.5 Results

Notes on participant actions and responses were recorded during the think-aloud tasks and in interviews. Notes consisted of brief comments made after notable responses to prompts, and for each interview question. The interviewer tagged participant remarks such as when they became stuck, noted confusion, indicated a preference, or commented on the interface. We used Dovetail for automatic transcription of our audio recording, and analyzed notes and transcriptions to extract larger themes. To do so, we counted the occurrences of describing words, preferences, and recurring utterances. We focused in particular on identifying usability issues, and to converge or choose from our two prototypes into a single design for our following study.

5.5.1 Variation Preferences. Most participants preferred the ANNOTATE design variation (Figure 5).

3 out of 8 participants found the information was more accessible with SUBTASKS [P2,P3,P7]. P4 liked the subtask structure as it was *"really satisfying to have, like all these subtasks that kind of led up to one bigger tasks"*. However, 3 also found it cluttered [P2,P4,P8]. While P2 preferred SUBTASKS, they pointed out that the interface could be potentially misleading because the parameters in SUBTASKS *"look like subtasks, but they're not, they're more like, changing the parameters of the tasks."* 4 participants described ANNOTATE interface as *"clearer"* and 2 as more *"integrated"*.

Variation preference varied between scenarios. Only P5 preferred the same variation (ANNOTATE) for all. For example, after stating a preference for SUBTASKS after CALL, when P7 finished TRIP they said they *"strongly preferred [ANNOTATE]"*: *"I liked the green underlined. It's growing on me. I've seen it now only three times. And it's very intuitive that it's clickable, that it's modular."*

6 participants stated both variations of the app were similar in terms of experience: *"They felt like very similar experiences. I'd say the strengths and weaknesses of the applications were the same across the two different variations. So I didn't really feel like the differences between the two variations had a strong impact on the experience."* [P1], *"Not a huge difference honestly it's a very small delta between the two"* [P7].

5.5.2 *Overall Interaction Paradigm.*

Participants were receptive to our approach: 5 of the 8 stated they would prefer using it over a VA. The remaining 3 claimed they were not familiar enough with VAs in general to make a comparison. Participants preferred our system for reasons that include speed [P2, P4, P8], ease of use [P1, P2, P8], and control [P4, P5]. *"This experience was good. I thought it was a lot quicker… It feels like a big time-saver and something that like really gets the job done"* [P4] (TRIP). Visibility was another reason for P1 and P2: *"I like to be able to look at all the information at once, whereas virtual assistants, like you just have to talk with them and keep a mental model of what you've already talked about."* [P1] (CALL). P4 also preferred the app to a virtual assistant because they felt they had *"a lot more ownership as to what's going on"*. In contrast, P5 had more difficulty navigating both prototypes, describing that they felt *"very rigid and difficult to wield. Like I'm fighting it, especially when I creating a new todo"*.

5.5.3 *Usability Issues.*

In general, participants were easily able to navigate both interfaces and to successfully complete their task. We did capture concerns and potential limitations. We identified three main usability issues with the initial variations of our prototype which we addressed before Study 2.

Inability feedback. 5 participants were curious or confused by the question mark icon in our original prototype whose purpose was to inform the user the underlying AI system does not understand and therefore cannot suggest an action. It was designed to disambiguate cases where the AI understand but is not able to act from cases where it cannot figure out user intent. P3 commented *"the whole question mark thing also confused me. I wasn't sure what that meant"*. We considered either adding a popup message explaining the icon's meaning, or removing it. We opted to remove it to reduce visual clutter, and disambiguating between inability to act and inability to understand was superfluous.

Subtask creation. 6 participants had some issue with the method to add a new task, specifically with how to create a new subtask. For example, P8 stated *"I was not too sure at first how I can add a new todo [SUBTASKS] under the task that I want"*, and P1 who thought *"creating new todos in the second one [ANNOTATE] was a little unintuitive by my personal tastes"*. After consulting existing todo app designs, we decided to add a stronger visual cue along with a prompt, to confirm and indicate the location of a new todo.

Task suggestions. 5 participants were confused or annoyed by the behaviour and appearance of the task suggestions: *"the suggestions are interesting, but I think the need to close off the other suggestions is a little annoying and can start to get out of hand"* [P5]. We decided to limit the number of task suggestions to keep the content focused on what the user's own tasks. We also rephrased tasks suggestions to be more relevant. These changes equates to having a stricter AI that has a higher threshold of what it think to be a relevant suggestion. We also considered indicating the AI's degree of certainty (similar to task parameter values) of relevancy of a task suggestion, but thought it would only add to the original issue of being too overwhelming.

5.5.4 *Prototype changes.*

We also made small improvements the both prototype variations and scenario prompts during the study based on observations and participant responses. We created a keyboard typing animation to make the prototype feel more realistic [P7], added more alternative user flows to make the prototype more interactive [P2], and resolve ambiguous phrasing in the scenario prompts [P3, P4, P6, P7]. These improvements led to our final prototype which we use in Study 2.

Most participants preferred the hypertext annotation variation. We selected this variation.

6 STUDY 2: COMPARISON WITH A VIRTUAL ASSISTANT PARADIGM

The goal of this second study is to compare our approach of non-anthropomorphic intelligent assistance against the widespread virtual assistant interaction paradigm. We created a final ToBeDone prototype refined from the findings of Study 1, and a virtual assistant (VA) prototype.

6.1 Participants

We analyzed the data of 13 participants, aged 19 to 22 (M = 20.8, SD = 1.0) recruited among peers. 10 participants self-reported as male, and 3 as female. 10 participants self-reported having written AI code or followed a course about AI, 2 having read about AI, and 1 reported knowing little about AI. 5 participants self-reported having used a virtual assistant at least once a week the past 30 days. Remuneration was a $10 gift card for a 1 hour study.

6.2 Task and Procedure

The task for Study 2 was identical to Study 1 but the procedure was modified, as the purpose of the second study was comparison instead of evaluation. For the interview after each scenario, we inquired about controllability ("What were your feelings of control?"), trust ("What were your feelings of trust?"), transparency ("Did you feel like you understood the capabilities of each interface?"), and preference ("How would you compare these two experiences?"). For interview at end of the session (after all 3 scenarios), the experimenter asked a single question ("Do you have any comments overall about these two experiences across the scenarios?").

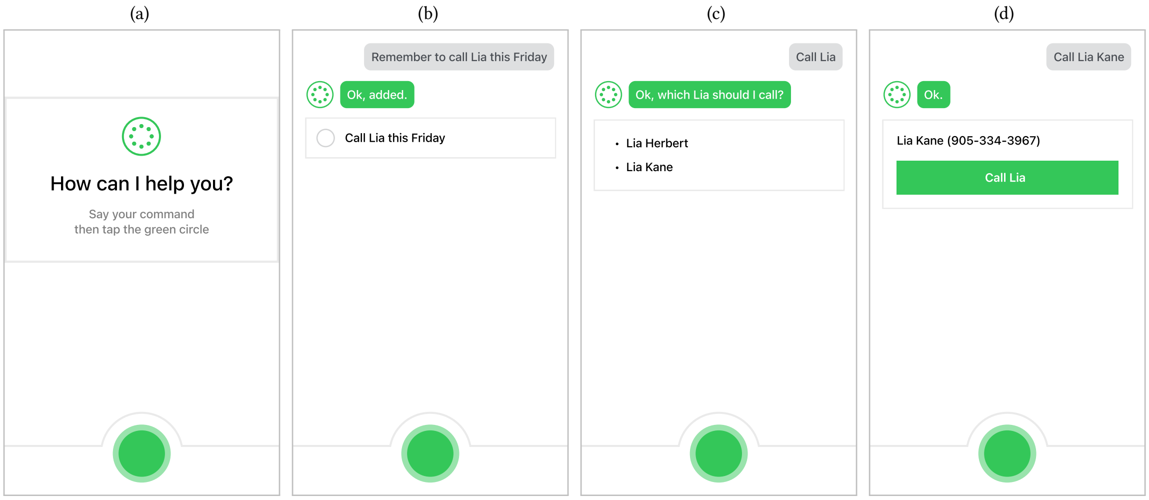
We also slightly adapted the scenarios from Study 1. We removed prompts that have the user remove a suggestion, for which the VA paradigm has no equivalent. We also shortened longer prompts to be more concise and less ambiguous, and used the exact terms for the user to specify, for example we changed "add ingredients" to specifically "red wine" and "chicken". This is because unlike ToBeDone, the VA paradigm is not able to guide the user with suggestions.

We provide videos demonstrating the study scenarios as supplementary materials.

6.3 Prototype

The two prototypes consisted of our final ToBeDone design and a virtual assistant (VA) interface. To enable comparability of the two paradigms, the capabilities and behaviour of the underlying AI system was identical, for example it made the exact same predictions and inaccuracies in both conditions.

Our VA interface shown in Figure 6 was inspired from representative popular commercial systems such as Apple Siri or Microsoft Cortana. Participants first spoke a command then pressed a button to submit it. They mostly interacted with the VA through speech, but confirmed commands and actions through touch. Our mockup simulated speech analysis and processing times for the same duration as ToBeDone. The prompts were identical, and the interface showed the same information, but in the style of virtual assistants. Both requests from users and responses from the underlying AI system were displayed as chat messages (Figure 6).



How can I help you?

Say your command
then tap the green circle

Remember to call Lia this Friday

Ok, added.

Call Lia this Friday

Call Lia

Ok, which Lia should I call?

- Lia Herbert
- Lia Kane

Call Lia Kane

Ok.

Lia Kane (905-334-3967)

Call Lia

Fig. 6. Virtual assistant paradigm on the CALL scenario in Study 2. (a) Welcome screen; (b) Recording a reminder; (c) Lifting ambiguity; (d) Making a call.

6.4 Design

This is a within subjects design with two paradigm conditions: TBD is our ToBeDone interaction paradigm, VA is the virtual assistant interaction paradigm. Each was evaluated in the context of 3 scenarios: CALL, MEAL, TRIP. The paradigm order was counterbalanced, scenario order was fixed according to expected increase in complexity.

6.5 Results

Results were analyzed in a similar manner as Study 1.

6.5.1 Controllability. Most participants found they had more control with the ToBeDone paradigm, in particular in more complex scenarios (Figure 7).

8 participants mentioned transparency as one of the main reason for feeling in control, qualifying the interface as *"straightforward"* [P9] and *"natural"* [P10]. P17 noted *"I preferred [TBD] because compared to [VA] I had more control. Specific things were underlined for me, I was able to go back and revise them and everything was on one page."* Some participants perceived the use of direct manipulation instead of an anthropomorphic assistant as the removal of an unwelcome intermediary: *"I feel more control when I can just input the tasks myself and edit it directly"* [P19], *"I felt I had more control in [TBD] because I could input stuff and change things without having someone do it for me"* [P18].

Some participants did reported feeling more in control with VA. The main reason was familiarity: 7 participants referred to similarity with existing virtual assistants. For example, P9 reported VA was *"pretty straightforward. Like I'm familiar with Siri and Alexa and Google Voice"*. But participants also mentioned VA felt more finished, closer to a commercial product [P17, P19] In contrast, 5 other participants reported lack of control when using the VA interface. For example, [P9] and [P11] mentioned *"it's hard to know what to say exactly, which is annoying"* [P11], and that it was too *"open-ended"* [P11].

4 participants initially found VA more controllable but changed their mind after the more complex MEAL and TRIP scenarios. P9 explained: *"As the task got more complex, it seemed like [VA] became less and less helpful"* [P9].

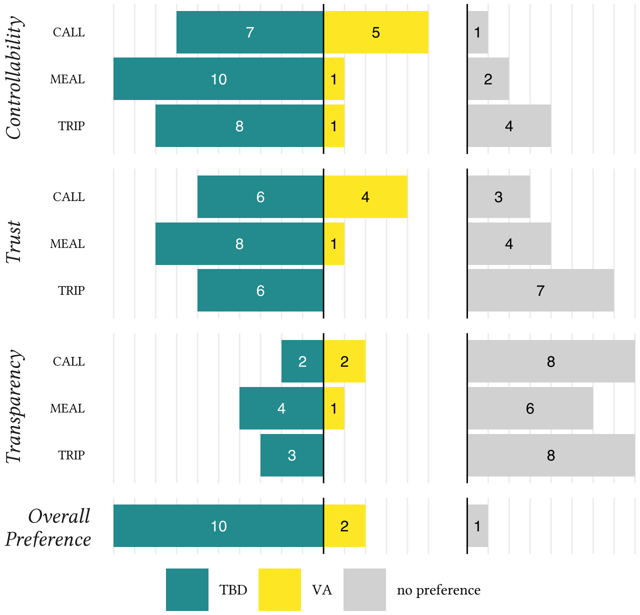


Fig. 7. Number of participants preferring each paradigm.

6.5.2 Trust. Most participants found ToBeDone more trustworthy than the VA (Figure 7).

In fact, participant dit not report any trust concerns related to the TBD paradigm. No participants exclusively distrusted it: only 3 said they distrusted it, and they also distrusted VA. They reported being hesitant to trust an AI to complete significant tasks, like booking a plane ticket.

In contrast, 10 out of 13 participants were dubious about VA speech interaction, and were worried the VA would misunderstand their commands: *"Verbally there's a lack of trust, because I'm less sure it will be interpreted correctly. With tapping things, it's very clear what I'm asking for."* [P10], *"Generally, I don't trust, voice all that much in general"* [P16], *"I didn't trust a speech recognition. If my goal was to call someone, somebody, I wouldn't trust a phone to recognize my voice and then call that person."* [P12] Yet, for comparability with the TBD condition that does not use speech as input, this type of inaccuracies was not included in the study.

6.5.3 Transparency. Overall, most participants reported they understood the capabilities of both paradigms in the post study questionnaire (Figure 7). Some participants did however gave a more nuanced response: P14 felt the VA interface was *"too magical"*, and P10 and P16 reported it felt like a *"black box"*.

6.5.4 Overall Preference. Overall, ToBeDone was greatly preferred (Figure 7).

2 out of 13 participants preferred VA while 10 preferred TBD, and only 1 had no overall preference. 5 participants stated they could see themselves using TBD in their everyday lives. P10 felt task suggestions made the interface *"a lot nicer to interact with"*, and P20 reported it makes it "easier to make edits". P18 found the VA *"a bit overwhelming"* and TBD *"less intrusive"*. However, P13 and P21 preferred the VA interface. P21 felt it provided a *"higher bandwidth communication between me and the system"* while TBD *"just seems like a lot of work"*. P13 liked that VA *"seemed more like a person"*.

7 DISCUSSION

Overall, participants responses to the ToBeDone paradigm were overwhelmingly positive. Trust, one of the main focus in virtual assistant research, did not seem to be a concern with it. Participants emphasized they found it easier to achieve complex tasks, and demonstrated a strong preference for ToBeDone.

7.1 Predictable and Controllable Assistance

As observed in previous work, many participants had trust concerns when using the VA paradigm. Speech input was one common issue, which may indicate concerns not about the paradigm itself, but rather about type of input that is put into use. It is likely rooted the participant's previous experience with current Voice Assistants because there was no voice recognition error in the study. Indeed, there is no equivalent with the ToBeDone paradigm and the system behaved the same way in both conditions.

Unlike speech input with which one often has to fully complete their request before knowing if it was properly recognized, text entry provides continuous feedback on user's input, and also enables granular error correction. This may be an other reason. It is possible that progress in speech recognition will help alleviate this issue in the future. Enabling users to fix a speech recognition result could enable them to benefit from faster speech input, and keyboard text-entry precision. It is worth noting that both type of inputs can be used with both paradigms, for example chatbots are a form of virtual assistant that relies on text input, and modern phone keyboards include a way to dictate text which could be used to enter a todo task.

When asked if they understood the capabilities of each paradigms, most participants responded affirmatively. However, our transparency question was asked after successful completion of the task, and may not properly reflect uncertainty before and during interaction. In fact, many participants did mention they were dubious of the system's ability to help them with their task, in particular with the VA paradigm which they qualified of *"blackbox"* [P10, P11] and *"ambiguous"* [P21]. VA interfaces provide little feedback about AI state and what information is used in action. In fact, 5 participants reported they thought the system would not be able to take context into account. When verbally formulating their requests, some participants also used a distrustful tone, indicating they did not believe the system would be able to help. On the other hand, the ToBeDone interface displayed all related information, showing users what inferences the AI made, and what to expect from the AI action. Very few participants reported being concerned by the ability of the AI system to help in the ToBeDone paradigm. We argue this indicate our emphasis on predictability as a design values was successful. We believe users do not have to trust the system because there is less uncertainty: the ToBeDone interface shows them what the AI is going to do before it does it. In fact, it even proactively does it so they do not have to consider if it is worth trying the AI.

7.2 AI Augmentation

Participants referred to the ToBeDone interface as *"straightforward"* [P9, P12], *"natural"* [P10], *"ordinary"* [P13, P21], as well as *"innovative"* [P11, P12]. The virtual assistant interface was perceived as *"cool"* [P9, P15], *"magical"* [P16], but also also *"unclear"* [P9, P15], *"confusing"* [P9, P15], or *"unnatural"* [P12, P21]. This indicates the success of our emphasis on augmentation rather than building a specific AI interaction tool like VAs. This was also reflected in participants intent of use, for example: *"[ToBeDone] was pretty standard. Something I would actually use like just solid right down and know of a to-do on my large to-do list. And then it was cool that I was able to integrate with phones in different ways. And someone's smart in that regard. [VA] I found like my initial experience was it, that was quite cool. Right? Like, it's cool to have this* *like voice powered virtual assistant, but I think the lack of trust I felt when it would mean that I wouldn't use that as a long-term solution."* [P9]

However, unlike the VA paradigm with which many participants were already familiar, participants did have to familiarize themselves with the style of using a todo to access AI systems. For example, P10 commented after completing the first scenario (CALL): *"[VA] was something I had done before so it was nothing unexpected. [ToBeDone] was just something I had never done before so it was more surprising."* This is intrinsic to any user interface relying on direct-manipulation principles [49], and participants were not provided with any help or explanations about the operations and feedback of the system. Yet, once they went through the first scenario, most were easily able to navigate and use the system. As an example, the same participant P10 later declared: *"[ToBeDone] felt very natural, it was much easier"* after completing the second scenario (DINNER).

Some participants reported finding virtual assistants appropriate for simple tasks, but that complexity is easier to manage with ToBeDone: *"[ToBeDone] was easier to follow than [VA] because compared to the [CALL] scenario, [DINNER] is more complicated and has several steps to complete the goal."* [P20], or *"As the task got more complex, it seemed like [VA] became less and less helpful"* [P9]. This may be explained by how ToBeDone feedback groups together relevant information, and inform the user what it will take into account. It is also possible to find the relevant information using a virtual assistant, for example by going through the VA conversation's history if available, or in some cases, by asking the VA. However, it is arguably more difficult than glancing the structured todo interface: *"with the virtual assistant, I can't get a visual of my tasks."* [P2], *"I like to be able to look at all the information at once, whereas virtual assistants, like you just have to talk with them and keep a mental model of what you've already talked about."* [P1], *"In [VA], the task felt too complex for me to communicate verbally at times with the computer, so I felt a lack of control there. [ToBeDone] felt very natural because it was easy to tap along."* [P10]. This tends to indicate that our emphasize on control and augmentation was successful: the system is easier to use in particular for complex task, and the AI overall more helpful.

We also believe one of the main advantages of augmentation is to always give users the option to complete the task manually. We did not investigate this particular aspect in our study because the VA paradigm does not incorporate a similar concept. Allowing users to ignore the AI seems more respectful and empowering, in fact previous work indicates users enjoy doing some things themselves [45, 49]. This also means that a task can likely still be achieved when the AI is inaccurate or incapable. This could be as simple as redirecting the user to a travel website to buy a ticket themselves, or even simply letting them record the task has been completed without any help from the system at all.

7.3 Limitations

There was an unavoidable legacy bias toward the VA paradigm. Indeed, virtually all our participants were familiar with the way VAs are operated, but none knew how to use our approach before the study: *"What I was trying to do was interact with it [VA] like Siri where I basically spoon-feed it all the information it needs in each request, because I assume it doesn't have the ability to keep context."* [P21] This was both detrimental and favourable for the VA condition. For example, in the simplest scenario, many participants found the VA approach easier. It is unclear if this is due to the direct nature of VA interaction or to existing familiarity with VA applications. However, as participants moved to more complex tasks, they expressed doubts about the VA's ability to accomplish a task, possibly because they previously experienced AI failures when using a VA. This biais can really only be alleviated with a longitudinal study and fully operational prototypes.

The three scenarios included in our study were carefully designed to cover a wide range of use cases including user intent refinement, document consultation, or prediction errors. However, they remain limited and necessarily misses many use cases. This work aims at demonstrating the value of the ToBeDone approach, and that it may be used to achieve most of what a virtual assistants do. Our results already hint that complex tasks tend to be easier with an ToBeDone paradigm, but more work is required to precisely identify the factors making a task easier to perform using one or the other paradigm.

8 CONCLUSION

This paper explores the design of general purpose AI system that augments direct interaction instead of replacing it, and help users in their every day tasks. The ToBeDone approach aims at providing an alternative to virtual assistants that is more transparent and controllable, and tightly integrated into users' workflow.

We chose a todo list application to demonstrate the possibility to design AI-augmented systems able to help users in their every day tasks. However, a similar approach can be applied to other form of applications, ranging from calendar to email and notepad, and future work is need to explore it. For example, a calendar application could learn from users routine, and suggest specific time slots to perform certain tasks like going to the grocery store, or a notepad application could suggest relevant resources such as website or academic articles related to what the user is writing. We hope this work will inspire researcher and designer to explore this path.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda.* Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. *Guidelines for Human-AI Interaction.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Denis Anson, Penni Moist, Mary Przywara, Heather Wells, Heather Saylor, and Hantz Maxime. 2006. The Effects of Word Completion and Word Prediction on Typing Rates Using On-Screen Keyboards. *Assistive Technology* 18, 2 (2006), 146–154. https://doi.org/10.1080/10400435.2006.10131913 PMID: 17236473.

[4] Michel Beaudouin-Lafon. 2004. Designing interaction, not interfaces. In *AVI '04 (AVI '04).* ACM, New York, NY, USA, 15–22. https://doi.org/10.1145/989863.989865

[5] Robbert-Jan Beun, Eveliene de Vos, and Cilia Witteman. 2003. Embodied Conversational Agents: Effects on Memory Performance and Anthropomorphisation. In *Intelligent Virtual Agents*, Thomas Rist, Ruth S Aylett, Daniel Ballin, and Jeff Rickel (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 315–319.

[6] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (jun 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[7] Ted Boren and Judith Ramey. 2000. Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication* 43, 3 (2000), 261–278. https://doi.org/10.1109/47.867942

[8] Scott Brave, Clifford Nass, and Kevin Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies* 62, 2 (2005), 161–178. https://doi.org/10.1016/j.ijhcs.2004.11.002

[9] Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine* 22, 4 (2001), 67–84. https://doi.org/10.1609/aimag.v22i4.1593

[10] Justine Cassell and Timothy Bickmore. 2000. External Manifestations of Trustworthiness in the Interface. *Commun. ACM* 43, 12 (dec 2000), 50–56. https://doi.org/10.1145/355112.355123

[11] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99).* Association for Computing Machinery, New York, NY, USA, 520–527. https://doi.org/10.1145/302979.303150

[12] Henriette Cramer, Vanessa Evers, Tim van Slooten, Mattijs Ghijsen, and Bob Wielinga. 2010. Trying Too Hard: Effects of Mobile Agents' (Inappropriate) Social Expressiveness on Trust, Affect and Compliance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10).* ACM, New York, NY, USA, 1471–1474. https://doi.org/10.1145/1753326.1753546

[13] Doris M Dehn and Susanne van Mulken. 2000. The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies* 52, 1 (2000), 1–22. https://doi.org/10.1006/ijhc.1999.0325

[14] Tom Djajadiningrat, Kees Overbeeke, and Stephan Wensveen. 2002. But How, Donald, Tell Us How? On the Creation of Meaning in Interaction Design through Feedforward and Inherent Feedback. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (London, England) *(DIS '02)*. Association for Computing Machinery, New York, NY, USA, 285–291. https://doi.org/10.1145/778712.778752

[15] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (jun 2003), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[16] Emir Efendić, Philippe P.F.M. Van de Calseyde, and Anthony M. Evans. 2020. Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes* 157 (2020), 103–114. https://doi.org/10.1016/j.obhdp.2020.01.008

[17] K. Anders Ericsson and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. The MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

[18] Junhong Gao and Michael S Brown. 2012. An interactive editing tool for correcting panoramas. In *SIGGRAPH Asia 2012 Technical Briefs*. 1–4.

[19] Kerstin Heuwinkel. 2013. Framing the Invisible – The Social Background of Trust. In *Your Virtual Butler: The Making-of*, Robert Trappl (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 16–26. https://doi.org/10.1007/978-3-642-37346-6_3

[20] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[21] Kate Hone. 2006. Empathic Agents to Reduce User Frustration: The Effects of Varying Agent Characteristics. *Interact. Comput.* 18, 2 (mar 2006), 227–245. https://doi.org/10.1016/j.intcom.2005.05.003

[22] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* 12, 4 (2000), 409–426. https://doi.org/10.1016/S0953-5438(99)00006-5

[23] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[24] Anthony Jameson, A Sears, and J Jacko. 2008. Adaptive user interfaces and agents. *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (2008), 433–458.

[25] Jonathan Klein, Youngme Moon, and Rosalind W Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with computers* 14, 2 (2002), 119–140.

[26] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[27] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics* 66 (2018), 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

[28] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan. 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[29] Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. ACM, New York, NY, USA, 264–275. https://doi.org/10.1145/2901790.2901842

[30] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[31] Ewa Luger and Abigail Sellen. 2016. *"Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[32] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, and Hongan Wang. 2019. *SmartEye: Assisting Instant Photo Taking via Integrating User Preference with Deep View Proposal Network*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300701

[33] Pattie Maes. 1994. Agents That Reduce Work and Information Overload. *Commun. ACM* 37, 7 (jul 1994), 30–40. https://doi.org/10.1145/176789.176792

[34] Pattie Maes and Robyn Kozierok. 1993. Learning Interface Agents. In *Proceedings of the Eleventh National Conference on Artificial Intelligence* (Washington, D.C.) *(AAAI'93)*. AAAI Press, 459–464.

[35] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[36] Don Norman. 2002. *The design of everyday things*. Basic books, New York.

[37] Donald A. Norman. 1994. How Might People Interact with Agents. *Commun. ACM* 37, 7 (July 1994), 68–71. https://doi.org/10.1145/176789.176796

[38] Erica L. Olmsted-Hawala, Elizabeth D. Murphy, Sam Hawala, and Kathleen T. Ashenfelter. 2010. *Think-Aloud Protocols: A Comparison of Three Think-Aloud Protocols for Use in Testing Data-Dissemination Web Sites for Usability*. Association for Computing Machinery, New York, NY, USA, 2381–2390. https://doi.org/10.1145/1753326.1753685

[39] Raja Parasuraman and Christopher A. Miller. 2004. Trust and Etiquette in High-criticality Automated Systems. *Commun. ACM* 47, 4 (apr 2004), 51–55. https://doi.org/10.1145/975817.975844

[40] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759. https://doi.org/10.1145/3379337.3415864

[41] Philip Quinn and Shumin Zhai. 2016. *A Cost-Benefit Study of Text Entry Suggestion Interaction*. Association for Computing Machinery, New York, NY, USA, 83–88. https://doi.org/10.1145/2858036.2858305

[42] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. *Explanations as Mechanisms for Supporting Algorithmic Transparency*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[44] Quentin Roy, Sébastien Berlioux, Géry Casiez, and Daniel Vogel. 2021. *Typing Efficiency and Suggestion Accuracy Influence the Benefits and Adoption of Word Suggestions*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445725

[45] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. *Automation Accuracy Is Good, but High Controllability May Be Better*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290605.3300750

[46] Nicole Shechtman and Leonard M. Horowitz. 2003. Media Inequality in Conversation: How People Behave Differently When Interacting with Computers and People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/642611.642661

[47] Thomas B Sheridan and William L Verplank. 1978. *Human and computer control of undersea teleoperators*. Technical Report. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

[48] Ben Shneiderman. 1987. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley Publ. Co.

[49] Ben Shneiderman. 1993. Beyond Intelligent Machines: Just Do It! *IEEE software* 10, 1 (1993), 100–103.

[50] Ben Shneiderman. 1995. Looking for the Bright Side of User Interface Agents. *interactions* 2, 1 (jan 1995), 13–15. https://doi.org/10.1145/208143.208150

[51] Ben Shneiderman. 1997. Direct Manipulation for Comprehensible, Predictable and Controllable User Interfaces. In *Proceedings of the 2nd International Conference on Intelligent User Interfaces* (Orlando, Florida, USA) *(IUI '97)*. Association for Computing Machinery, New York, NY, USA, 33–39. https://doi.org/10.1145/238218.238281

[52] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[53] Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. *Interactions* 4, 6 (Nov. 1997), 42–61. https://doi.org/10.1145/267505.267514

[54] Barry Strauch. 2018. Ironies of Automation: Still Unresolved After All These Years. *IEEE Transactions on Human-Machine Systems* 48, 5 (Oct 2018), 419–433. https://doi.org/10.1109/THMS.2017.2732506

[55] Jo Vermeulen, Kris Luyten, Elise van den Hoven, and Karin Coninx. 2013. *Crossing the Bridge over Norman's Gulf of Execution: Revealing Feedforward's True Identity*. Association for Computing Machinery, New York, NY, USA, 1931–1940. https://doi.org/10.1145/2470654.2466255

[56] Janet H Walker, Lee Sproull, and R Subramani. 1994. Using a Human Face in an Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 85–91. https://doi.org/10.1145/191666.191708

[57] Joseph Weizenbaum. 1966. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168

[58] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468