Kindly 2.0: Human-in-the-Loop AI for Conversational Agents

PART 1: Innovation

1. Underlying idea

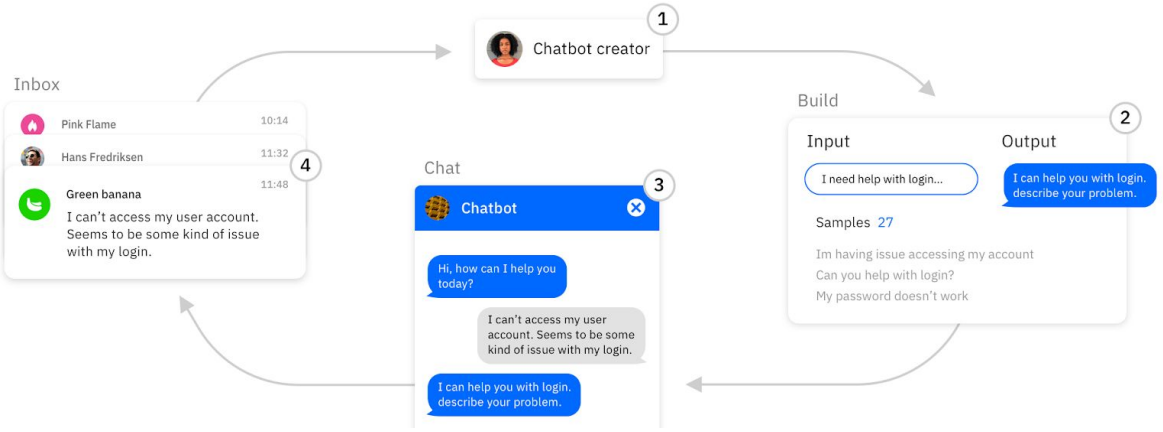

Figure 1: In chatbot creation platforms, humans (1) define intents through lists of user samples (2) which are then used as training data to reply to end users (3). Chatbot creators then further develop a system based on the observed chatbot interactions (4).

In recent years, chatbots have become a ubiquitous technology in customer support, enabling companies to provide round-the-clock automated assistance to their customers. By solving simpler customer support queries that do not require human assistance, they allow support professionals to focus on the more demanding cases, drastically increasing the volume of solved customer support inquiries.

Many existing chatbot are based on ad-hoc software solutions developed and maintained by in-house or contract engineers. This approach, however, is costly in terms of time and resources, and becomes increasingly difficult to maintain as the system grows in complexity. The Kindly platform takes a radically different approach and is designed from the ground up to enable non-technical users (the support professionals, from now on referred to as "chatbot creators") to build and deploy their chatbot. By combining state-of-the art Machine Learning (ML) and Natural Language Processing (NLP) techniques with a strong focus on ease of use, Kindly has become a leading product in the Scandinavian market, being successfully adopted by industry-leading companies such as Adecco, Norwegian, Elkjøp, Finn.no and Thon Hotels, as well as smaller enterprises such as Sex og samfunn, Huseiernes Landsforbund and Kommuneforlaget, to name a few.

The role of chatbot creators in Kindly is summarized in Figure 1 above. They start by creating the training data used to learn the AI model powering the chatbot (steps 2 and 3, corresponding to the so-called *Build* section of Kindly). The chatbot creator can then observe the interactions with the end users in the conversation logs (the *Inbox* section in Kindly, step 4), and further improve the chatbot by modifying the training data based on these observations. While Kindly and similar platforms do make chatbot creation and deployment practical and relatively seamless for their users, several aspects remain quite challenging. In order to train the Natural Language Understanding (NLU) engine of their chatbot, Kindly users attempt to capture the *intents* of prospective chatbot users by writing a set of examples ("How do I change my password", "I lost my password", "Where do I go to update my password?", and so on) that map to a

common answer ("To change your password, go to www.mysite.com/reset-password"). An NLU model, typically a deep neural network with potentially millions of parameters, is then trained behind the scenes, allowing the chatbot to answer queries that are semantically similar to the ones in the training data.

To improve the performance of the chatbot, Kindly users currently need to actively monitor the chatbot logs, observe how end users interact with the chatbot, and increase the coverage of a given intent with more examples, remove (or edit) examples that might have led to a wrong answer, or create new intents altogether. This approach is problematic on two fronts: (a) it fails to leverage the information contained within the chatbot interactions directly, since Kindly users need to manually inspect the conversation logs and find out how to edit the training examples to account for the observed errors, and (b) they have to attempt to guess what led the NLU algorithm to respond in a certain way. Another challenge in Kindly (and similar platforms) is that (c) dialogue flows are restricted to specific paths, where the user interacting with the chatbot is expected to always stick through the same topic and answer the chatbot questions in a specific way. This causes problems when the user diverges from this predefined flow; for instance, when they want to switch to another topic or formulate their answers in a different way than initially envisioned.

The Kindly 2.0 project intends to tackle these technological challenges by taking advantage of recent developments in NLP and machine learning research. The project's key idea is to rely on a **human-in-the-loop AI**[1] approach where the human chatbot creators and the data-driven AI models work hand-in-hand to continuously enhance the conversational capabilities of the chatbot.

In particular, this project will deliver:

(a) **Semi-automatic data annotation**, which has the potential of drastically reducing the effort to train the intent recognition model of a chatbot, while improving accuracy;
(b) **Transparent NLU models**, which will allow chatbot creators to gain a better understanding of the NLU model decisions, making it easier to take steps to further improve their chatbot;
(c) **Context-aware dialogue management**, which will enable chatbot creators to design less rigid conversation flows by taking into account the current dialogue context (such as the previous user inputs) to determine the most appropriate response.

The combination of Kindly data science team's expertise in bringing state-of-the-art NLP systems to a framework for non-technical users and the Norwegian Computing Center's (NR) excellent track record in NLP research, along with our partners' expertise in using chatbot creation platforms puts us in the unique position to develop and roll out the next generation of chatbot development platforms.

2. Level of innovation

The proposed innovation will empower chatbot creators with new functionalities for building, deploying and monitoring chatbots without the need of advanced technical expertise. Concretely, the planned R&D efforts will result in substantial improvements in Kindly's *Build* and *Inbox* sections, upgrading the quality of chatbots both from a creator and an end user perspective, leading to the radically revamped version of the Kindly platform: **Kindly 2.0**. The new, AI-driven software solutions that will arise from the project are expected to not only benefit the project participants, but also Kindly customers at large.

The project is also likely to positively contribute to the language technology sector in Norway. In particular, since this project's commercial partners operate their chatbots in Norwegian, the project is also expected to

[1] "*Human-in-the-loop AI*" refers to a family of machine learning techniques that are specifically designed to take advantage of human expertise in the development and evaluation of data-driven models.

boost the creation of language resources for Norwegian, which remain relatively scarce compared to other languages. Furthermore, the development of innovative techniques for weak supervision and explainable models is likely to be useful to many other Norwegian AI-driven companies.

If successful, the innovative processes for AI training developed in this project will radically improve the way Kindly users create and maintain their chatbots, improving not only the quality of end-user conversation, but also the day-to-day work of chatbot creators. Concretely, Kindly 2.0 will allow them to create robust chatbots with modest amounts of training data (making the task less repetitive), and increase their understanding of the decisions taken by the AI algorithms in the platform (mitigating the frustrations that can originate from working with a black-box system).

The innovative technological solutions developed in this project will create new business opportunities for Kindly, notably by attracting customers that do not have enough resources to produce large training data for chatbots. These solutions are not only beneficial for Kindly but also for small and mid-sized enterprises as it lowers the threshold of adopting AI technologies in their customer support departments.

3. Potential for value creation

To the extent of our knowledge, there is no existing commercial chatbot platform that offers functionalities similar to the ones that will be developed through this project. The technological innovations put forward by this project will thus substantially improve Kindly's market position with regard to competing platforms.

The potential value creation for Kindly is a larger market share in the scandinavian market where Kindly already has become a leading product in the chatbot space. In addition to the scandinavian market, there is a larger potential in expanding globally to markets where the company currently does not operate. The ambitions of Kindly is to become a leading chatbot platform globally. The new functionalities described in this proposal will allow Kindly to further expand into this relatively new and innovative market where startups are able to compete with large enterprises such as Google, Microsoft, Facebook. Amazon, and Nuance. In a recently-published report, Zion Market Research estimates that the global chatbot market could be evaluated to around USD 369.79 million in 2017, and that it is expected to grow to USD 2,166.28 million in 2024 (Zion Market Research, 2018). The potential value creation is vast, if Kindly manages to scale its success and operations outside the scandinavian market and towards the much larger European, American and global markets.

Kommuneforlaget's core mission is to improve the quality of public services through digital solutions and educational technology. Kommuneforlaget started to work with Kindly on a digital solution for municipalities in Norway in 2018, and so far delivered chatbot solutions to 25 municipalities in Norway, with the aim of scaling this solution to as many municipalities as possible. The potential value creation for Kommuneforlaget is the ability to deliver better chatbot solutions to municipalities.

For Norwegian Airlines AS and Adecco Norway, the potential for value creation is the early access to functionality by being part of this project. With the potential success of the hypothesis laid forth in this research and development project, there is an even larger potential for value creation for these two companies, and all of the Kindly's customers, to create much more precise and complex chatbots that will handle ever larger part of their user and customer dialogue.

4. Project participants and constellation of partners

4.1. Research-performing and financing partners

a) Companies in Norway that will be using the project's R&D results in their own value creation.

C1: Convertelligence AS (Project owner)
Convertelligence is a technology company based in Oslo (Norway) founded in 2016 by Arash Saidi and John Antonio Nilsen, whom are still the major owners and co-CEOs of the company. By year-end 2018, the company counted 40 employees (recalculated to 34 FTE) with revenues of NOK 11.3 millions compared to NOK 4.1 millions in 2017 (170% growth). Convertelligence is the sole owner of the subsidiary Kindly AS.

The Kindly platform provides intelligent chatbots for some of the largest enterprises in the Nordics, such as Norwegian Air Shuttle, Elkjøp, Kahoot!, Thon Hotels, Statnett, Visma and Finn. Kindly employs a data science team with a strong track record of internationally published NLP research. Previous work of our data scientists has brought significant contributions to NLP sub-fields that are crucial for the development of robust NLU systems, including: document classification, semantic analysis, linguistic representation, syntactico-semantic parsing, NLP infrastructure, and deep learning methods for NLP.

C2: KF (Kommuneforlaget)
KF has been named the 6th most innovative company in Norway, by technology magazine INNOMAG (Norway's leading independent digital innovation magazine). Their goal is to improve the quality of public services through digital solutions and educational technology. All 422 municipal authorities subscribe to one or more of their digital products. They have digital products within four key areas: 1) Digitalization, 2) Quality and internal control, 3) Data collection and analysis, and 4) Adolescence and education.

C3: Norwegian Airlines ASA
Norwegian is the fifth largest low-cost carrier in the world with around 11,000 dedicated colleagues and a modern, environmentally friendly fleet of around 170 aircrafts with an average age of 3.8 years. Norwegian believes in affordable fares for all and their customers can choose from around 500 routes to more than 150 destinations in Europe, North Africa, the Middle East, Asia, the Caribbean, South America, and the US. In 2018, Norwegian carried more than 37 million passengers and has been voted 'Europe's best low-cost carrier' by passengers for six consecutive years at Skytrax World Airline Awards (2013-2018), as well as the 'World's best low-cost long-haul airline' for five consecutive years (2015-2019).

C4: Adecco Norway
The world's leading workforce solutions company, offering temporary staffing, permanent placement and outsourcing across all sectors. Having offices in 62 locations all across Norway gives them a great insight and knowledge of the local markets. Cooperation between different areas and fields of expertise allows them to create solutions fast and efficiently. Adecco has around 400 employees in Norway. They have about 6000 recruited workers out in the field every day and receives approximately 1000 applications every week. Adecco Norway is a part of the global company The Adecco Group.

By participating in this project, all three partner companies intend to improve their chatbot creation and development processes as well as inform and influence the general R&D direction of the Kindly platform. Since all partners are professional users of the Kindly platform, they will also be able to provide vital feedback in the process; notably by validating early-to-late versions of the models and techniques developed throughout the project. Coming from three distinct business domains, these partner companies ensure a diversity not only in the type of dialogues they build, but also in the type of queries their chatbots are

designed to answer and the number of end-users (Norwegian being the one with the most traffic). They will therefore provide Kindly with valuable feedback that will be relevant to customers across many fields and business domains. By participating in this project, the partners will also be able to get first-hand knowledge of the developments in the platform, as well as gain early access to new features. Each of the companies will provide personnel to participate in testing the developments throughout the project period.

b) Partners from the R&D sector

R1: Norwegian Computing Center
The Norwegian Computing Center (NR) is one of Norway's leading research institutions within statistical modelling, machine learning and ICT. Established in 1952, NR is organised as a private, nonprofit foundation that carries out R&D projects for a broad range of commercial and public organisations in Norway and internationally. NR leads the BigInsight centre of excellence for research-based innovation, which aims to produce innovative solutions for the knowledge economy through novel statistical and machine-learning methodologies for extracting actionable knowledge from complex data. NR has about 85 employees, a majority of whom are research scientists with a Ph.D.

4.2. Other forms of collaboration

The project will have an advisory board consisting of *Erik Velldal* (Associate Professor at the University of Oslo), *Fredrik Jørgensen* (Staff Data Scientist at Schibsted) and *Bente Kvam Kristoffersen* (Managing Director & Partner at Trigger Oslo). The advisory board will provide feedback on the project, in particular on academic, professional and commercial dissemination of the outcomes of the project. Further, Convertelligence will co-supervise MSc projects with the Language Technology Group at the University of Oslo.

PART 2: The R&D activities

5. Need for research

Most chatbot platforms currently on the market rely either on handcrafted rules or machine learning techniques (or a combination of both). However, both families of approaches have important shortcomings:

1) *Rule-based approaches* often have substantial upfront development costs due to the need to write, validate and maintain a large number of dialogue rules. Furthermore, they tend to lead to relatively rigid dialogue flows in which the user is forced to interact with the system in a specific manner and cannot deviate from this prespecified flow (for instance if they wish to answer the chatbot in a different way than prescribed by the rules).
2) *Machine learning approaches*, on the other hand, are often less costly to deploy, but are dependent on large amounts of in-domain training data. Unfortunately, in most domains, such training data is either available in small quantities only or not available upfront. Furthermore, the behaviour of such data-driven chatbots is often difficult to interpret (e.g. neural network-based chatbots are constructed out of millions of numerical parameters), and even harder to modify.

This project intends to bridge the gap between these two families of approaches by developing chatbot models that can leverage both expert domain knowledge and machine learning in order to "get the best of both worlds". In particular, the project will rely on a *human-in-the-loop* AI approach, seeking to take

advantage of the expertise acquired by domain experts through the use of *weak supervision* and *explainable* machine learning models.

Weak supervision (Ratner et al. 2017) is a relatively recent AI paradigm that allows machine learning models to be learned from a combination of noisy supervision signals instead of large amounts of manually labelled data. Crucially, these supervision signals may take the form of heuristics written by domain experts. Although weak supervision has attracted a lot of attention in academia, only a handful of studies (Mallinar et al. 2019) have explored its use for the development of conversational agents. Furthermore, using weak supervision in real-life application remains a largely uncharted territory with a few very recent exceptions such as the collaboration between Google and Stanford University in (Bach et al. 2019).

Similarly, although the field of explainable ML has grown in recent years, the development of explainable conversational models remains in its infancy (Galitsky and Goldberg 2019), limited to theoretical studies which are—first and foremost—concerned with "explainability" for ML experts. The lack of research on *explainability for non-technical users* is a significant obstacle in chatbot development, as chatbot creators attach great importance to understanding and controlling how their chatbot interacts with end users (since the chatbot represents the outward face of the company to their customers).

6. Objectives

The primary objective of this project is to design, implement and validate a new software framework for developing high-quality conversational agents. Using a range of innovative AI techniques, Kindly 2.0 will make it easier for Kindly's customers to build a chatbot for their application and enhance its behaviour over time. This primary objective will be achieved by pursuing three subsidiary goals:

1) Develop data-driven models for intent recognition and response selection using multiple weak supervision sources such as expert heuristics and dialogue annotations.
2) Develop new dialogue management models to tackle complex NLU scenarios and allow for more fluid dialogue flows.
3) Develop a dialogue analysis tool to empower chatbot creators by allowing them to understand and analyse the chatbot behaviour at multiple levels of granularity (from single responses to global quality metrics), annotate the responses, and leverage such annotations to continuously improve the chatbot.

7. R&D challenges and scientific methods

The project will address three important limitations in the current state-of-the-art for chatbot development:

1) **Scarcity of training data**. In most practical settings, chatbot developers only have access to small amounts of annotated dialogue examples (if any). This causes problems when the chatbot encounters user inputs that were not covered in the initial training set.
2) **Rigid conversations**. Most chatbots are quite limited in the way they allow end users to interact with them. However, the behaviour of end users is often difficult to predict and may deviate from this predefined flow. They may for instance the system prompts in a manner that was not anticipated by the system, or even decide to switch topic in the middle of a conversation.
3) **Explainability.** The explainability of deep learning models have become a pressing issue in the field. Often referred to as a "black box", a neural network contains a multitude of parameters which are almost impossible for humans to understand. Recently, there has been a surge in the number of studies that try to peek into the "black box" and understand its behaviour. However, this thread of research is still in its infancy and, more importantly, it has been almost exclusively focused on explainability for experts (i.e.

researchers). There remains a gap to be filled when it comes to explainability, in general, and for the wider public in particular, i.e. non-technical users of ML such as chatbot creators.

The project sets out to address these three challenges using several novel ML techniques, described below.

Scarcity of training data

Weak supervision (Ratner et al. 2017; Mallinar et al. 2019) will be used to expand the size of the initial training data by leveraging a wider variety of (possibly noisy) supervision signals. These signals will be generated from multiple sources such as:

1) Heuristic functions (such as surface patterns) written by domain experts (i.e. chatbot creators)
2) Models trained on the same task but on a slightly different domain (thereby allowing the knowledge acquired from one chatbot domain to be transferred to other domains)
3) Annotations of existing dialogue data (viz. marking chatbot responses in the chatlog as correct or incorrect - and in this case providing a more appropriate response)
4) Expansions of a small initial set of manually constructed examples (for instance by replacing some of the words in these examples by synonyms)

This weak supervision strategy will be applied to the two central components of chatbots, namely: *intent recognition* (which is responsible for inferring the most likely user intent based on the observed user inputs) and *response selection* (which is responsible for selecting the most appropriate response given the inferred intent and the surrounding context).[2] A range of weak supervision sources will be defined for each of these two tasks, and then employed to estimate a probabilistic label model that accounts for the fact that supervision sources may have varying accuracies as well as correlations between them (Ratner et al. 2019).

Rigid conversation flows

The project will develop more flexible strategies for working with task-oriented dialogues, aiming to make the conversational agents more "context aware" (Tian et al. 2017) and able to conduct complex interactions with end users. This context-awareness requires the ability to (1) track the current state of the dialogue over time and (2) select different responses to the user depending on this dialogue state.

Take the following fictive dialogue as an example:

(1) **System:** *Hi, how can I help you?*
(2) **User:** I would like to book a flight
(3) **System** *Sure, what is your destination?*
(4) **User:** Hamburg
(5) **System:** *OK, Hamburg. And what is your departure?*
(6) **User:** I wish to leave from Oslo on November 3
(7) **System:** *We have a flight from Oslo to Hamburg departing at 8:05 AM on November 3, at a cost of 2500 kr. for a one-way trip. Would you like me to book it?*
(8) **User:** You know what, nevermind Hamburg - what are the flight options to Bremen instead?
(9) **System:** *No problem. Do you still want to leave from Oslo on November 3?*

[2] For instance, given a user input such as "*I would like to cancel my flight booking, how do I do that?*", intent recognition is responsible for mapping the sentence to the intent *Cancel(FlightBooking),* while response selection will map this intent to the response *"I can take care of cancelling your booking. What is your reservation number?"*.

Although superficially simple, the dialogue above exhibits conversational phenomena that are difficult to handle with existing chatbot technology. First, the user response at (6) contains more information than expected (the system only asked for the departure, but the user provided both the departure and departure date). Many commercial chatbots are designed such that they expect the user to only answer the current question and nothing else, and are thus likely to respond with the follow-up question: "*And what is your departure date?*", ignoring the fact that this information was already provided by the user. Second, the user response at (8) indicates that the user wants to drop their current request and start a new one. Most chatbots are hard pressed to handle this type of topic-switching in the middle of a dialogue, and will often require the user to fulfill their current request until the end (or abort the conversation altogether).

The project will develop robust models for both dialogue state tracking and response selection, taking advantage of recent advances in the field based on e.g. transfer learning and multi-task learning (Gao, Galley, and Li 2019; Rastogi, Gupta, and Hakkani-Tur 2018), in addition to weak supervision techniques.

Explainability

The project will also develop new methods for improving the **explainability** of the chatbot decisions. One major impediment for the deployment of machine learning models in commercial systems stems from the fact that these models are hard to understand, even for machine learning experts. In the past few years, the question of explainability has become a very active research topic in machine learning and NLP, and various solutions have emerged. Local explanation methods such as LIME (Ribeiro, Singh, and Guestrin 2016) can be employed to determine which words or expressions were most influential to determine the most likely intent of a given sentence. Other approaches such as (Koh and Liang 2017) look for samples in the original training data that can best explain the model decision for a given input. However, recent results have shown that explainability methods such as attention mechanisms are not always reliable (Jain and Wallace 2019) and more research is needed to determine their applicability to different tasks, not least in the context of ML-based conversational agents. In addition, the project will also look at global explanations methods that can "distill" the knowledge of a deep learning model into a more interpretable model such as a decision tree (Frosst and Hinton 2017) which have an intuitive visual representation.

Finally, as mentioned above, a key focus of the R&D efforts will be to achieve a explainability level that is also understood by chatbot creators who do *not* have a background in ML, or AI in general.

Development challenges

Implementing all of the components and ML techniques described above will require substantial development of the Kindly chatbot platform—both on the frontend (viz. UX and UI design) and backend levels. In addition to the engineering aspect, such development efforts will need to tackle, among other things, the question of how to present the new AI techniques to chatbot creators in a simplified yet accurate manner that enables them to leverage the potential benefits of cutting-edge ML techniques.

Even though different bits and pieces will require refactoring in the Kindly platform, we focus here on two central components that will need substantial development, namely: *Build* and *Inbox* (see Figure 1 in §1). Weak supervision will be implemented under the *Build* component which is used by chatbot creators to, among other things, define dialogue flows and their corresponding training data. Introducing weak supervision to *Build* requires developing a functionality that allows chatbot creators to programmatically define labelling functions that operate on the chat logs and extracts weak supervision signals. Likewise, implementing flexible dialogues relies on the development of *Build* to allow chatbot creators to define task-oriented flows with intuitive and effective control on the tracking of dialogue states.

The explainable NLU models will be implemented under the *dialogue analysis tool*, which will serve as an extension of Kindly's *Inbox*. The dialogue analysis tool will provide several functionalities such as:

- The ability to analyse each chatbot response in detail (using the explainable models described above), and inspect the reasoning pattern that led the model to infer a particular user intent.
- Evaluation metrics to assess the overall quality of the interaction. The choice of evaluation metrics is a non-trivial problem, as there is no easy way to determine what makes a dialogue successful (Tian et al. 2017; Liu et al. 2016). The project will investigate a spectrum of possible metrics, select the ones most appropriate for each application, and present these measures in a user-friendly interface (making it possible to e.g. easily search for good or problematic interactions).
- The analysis tool will also go beyond mere reporting and develop methods to automatically discover miscommunication patterns in the conversation logs, for instance, user requests that were not understood by the system. More specifically, the project will implement outlier detection models to automatically detect miscommunication patterns in conversation logs (Meena et al. 2015).
- Finally, chatbot creators will be able to automatically annotate the dialogues, by flagging some chatbot responses as correct or incorrect. These annotations will be subsequently exploited as weak supervision source for data-driven models for intent recognition and response/action selection.

Thanks to these novel functionalities for chatbot development and analysis, the behaviour of the chatbot will no longer be settled once and for all when the chatbot starts to be deployed. In other words, by leveraging the content generated by the end users, chatbots can be continuously refined, improved and made more robust to unforeseen dialogue scenarios as more dialogues are collected, evaluated and annotated.

8. Project plan

8 a) Main activities ("work packages") under the project

WP0 Project management	**WP Type** N/A
Participants Convertelligence, all	
Coordination of R&D activities, administration and internal communication, organisation of project-related events, reporting, and quality assurance.	

WP1 Weak supervision for intent recognition	**WP Type** Industrial Research
Participants Convertelligence, NR	
Goal: Develop data-driven models for intent recognition based on multiple weak supervision sources. This WP (and WP2 and WP3) will follow an empirical methodology with iterative cycles of development and evaluation on manually annotated datasets (cf. WP5). **Deliverables**: - Prototype of intent recognition models with weak supervision - Submission of papers on weak supervision for intent recognition to relevant ML/NLP conferences - Public seminar hosted by Convertelligence to share the results with ML and NLP community in Norway	

WP2 Flexible goal-oriented dialogues	**WP Type** Industrial Research
Participants Convertelligence, NR	

Goal: Develop a new data-driven modeling approach for response selection in goal-oriented dialogues, making use of the conversation history.

Deliverables:
- Prototype of goal-oriented dialogue system
- Submission of a research paper to the SIGDial conference

WP3 Explainability for ML models	**WP Type** Industrial Research
Participants Convertelligence, NR	

Goal: Techniques to derive human-understable explanations of the ML models powering Kindly chatbots.

Deliverables:
- Prototype of the methods to explain the chatbot's 'decisions' for intent and response selection
- Publication in relevant NLP conferences such as the BlackboxNLP workshop

WP4 Dialogue analysis and control	**WP Type** Industrial Research
Participants Convertelligence, NR	

Goal: Develop new techniques to assist chatbot creators in analysing and annotating chatbot conversations, evaluating the chatbot behaviour at multiple levels of granularity: from the level of individual responses to the detection of miscommunication patterns and the use of dialogue-level evaluation metrics.

Deliverables:
- Prototype software tool, including evaluation metrics and outlier detection model

WP5 Data annotation and evaluation	**WP Type** Experimental development
Participants Convertelligence, Norwegian, Adecco, Kommuneforlaget	

Goal: Collect manually annotated (gold standard) datasets to conduct methodological evaluations of the models developed in WP1 and WP2. This WP is different from WP7 in that it aims to achieve scientifically sound evaluation of the new models in a controlled environment while using real-life examples.

Deliverables:
- Manually annotated datasets for weakly supervised intent recognition and goal-oriented dialogues
- Paper submission to a scientific conference (such as NoDaLiDa) in 2021.

WP6 Implementation and infrastructure development	**WP Type** Experimental development
Participants Convertelligence	

Goal: Backend and frontend development and user experience (UX) R&D of Kindly's platform to integrate the new ML techniques in WPs 1, 2, 3 and 4 in the Kindly platform.

In addition to integrating the prototypes from WP1-WP5 in a production environment with high quality standards, the work package will also investigate how to make such new functionalities "accessible" to chatbot creators (who have no knowledge of ML and NLP). The latter point—in and of itself—remains an open R&D question that will involve researchers, developers, UX designers, ML engineers as well as end-users from partner companies such as Norwegian and Kommuneforlaget.

Deliverables: - Kindly 2.0 chatbot creation platform - Workshop with Kindly's customers to introduce the new features and ML techniques in Kindly 2.0

WP7 Pilot experiments and validation	**WP Type** Experimental development
Participants Convertelligence, Norwegian, Adecco, Kommuneforlaget	
Goal: Validate the results of WPs 1, 2, 3, 4 and 6 by partner companies in a series of pilot experiments. **Deliverables**: - Feedback summary for the final implementation of the new features in the Kindly platform.	

The duration of the work packages above is detailed in the table below.

	2020				2021				2022
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
WP0									
WP1									
WP2									
WP3									
WP4									
WP5									
WP6									
WP7									

8 b) Budgeted project costs distributed by Main Activity

Table 8 b)

No.	*Title of main activity / work package*	*Budgeted costs: (NOK 1000)*	*Costs: Industrial research*	*Costs: Experimental development*
WP0	Project management	-	-	-
WP1	Weak supervision for intent recognition	5 525	5 525	-
WP2	Flexible goal-oriented dialogues	5 525	5 525	-
WP3	Explainability for ML models	5 525	5 525	-
WP4	Dialogue analysis & control	5 525	5 525	-
WP5	Data annotation and evaluation	3 122,5	-	3 122,5
WP6	Implementation & infrastructure	6 145	-	6 145
WP7	Pilot experiments and validation	1 632,5	-	1 632,5
Total	***Entire project***	**33 000**	**22 100**	**10 900**

8 c) Critical milestones for the R&D activities

Milestone name	**Related WPs**	**Estimated date**	**Description**
First prototype version	1, 3,5,6	2021/Q1	First version of Kindly 2.0, integrating weak supervision sources and explainable model decisions.
Second prototype version	2,3,4,5,6,7	2021/Q4	Second version of Kindly 2.0, including the creation of flexible, goal-oriented dialogues and the dialogue analysis tool. Improved functionalities for weak supervision and explainable chatbot models.
Stable version	5,6,7	2022/Q1	Integration of all components and final validation

8 d) Project organisation and management

Convertelligence will be responsible for all project activities, with substantial contributions from the other project partners. In particular, NR will be tightly involved in all research-driven project activities (WP1-WP4) along with scholarly dissemination, while Norwegian, Adecco and Kommuneforlaget will contribute to the evaluation and validation activities in WP5-WP7.

The project will be led by *Emanuele Lapponi*, who currently holds the position of Lead Data Scientist at Convertelligence. Lapponi has an extensive R&D experience with language technology and machine learning, and has recently submitted his PhD in Computer Science at the University of Oslo. Due to his familiarity with both the research and commercial sectors, he is ideally suited to lead the proposed project.

Pierre Lison will lead NR's contribution to the project. He has a long research experience in NLP and in particular in data-driven dialogue modelling, and has led several R&D projects in this field. As mentioned in §4, the project will also rely on an advisory board, consisting of Erik Velldal (UiO), Fredrik Jørgensen (Schibsted) and Bente Kvam Kristoffersen (Trigger Oslo), to provide regular feedback on the project progress and future directions.

Table 8 d) Distribution of tasks and responsibility in the project

Partner	***Name of partner***	***Responsible for main activity:***	***Participating in the following main activities:***
C1	Convertelligence	WP0-WP7	WP0-WP7
C2	KF		WP5, WP6, WP7
C3	Adecco		WP5, WP6, WP7
C4	Norwegian		WP5, WP6, WP7
R1	NR		WP0, WP1, WP2, WP3, WP4

9. Funding

Table 9. Distribution of funding

Partner	*Name of partner*	*Own financing (NOK 1000)*	*Other funding (NOK 1000)*	*Total (NOK 1000)*
C1	Convertelligence	30 220		30 220
C2	KF	60	-	60
C3	Adecco	60	-	60
C4	Norwegian	60	-	60
Overall funding from project partners		**30 400**	**-**	**30 400**
Amount sought from the Research Council				16 000
Total funding (= total project costs)				33 000

PART 3: Plan for implementation and utilisation of results

10. Implementation plan for value creation for the company partners

The implementation plan for value creation for Convertelligence as the project owner is to incorporate the research and development outlined in the work packages (see §8) as new features and functionalities in the Kindly chatbot platform. Kindly's sales and marketing teams will work closely with the technical team in order to create marketing and sales strategies for the R&D outcomes delivered by the project.

The chatbot market is a highly competitive space with many startups and large enterprises attempting to solve the challenge of automating the highly complex task of human conversations. In the Scandinavian markets there are two main competitors to Kindly's chatbot platform: Boost.ai and Artificial Solutions. Both competitors deliver chatbot solutions to large enterprises. Companies that choose Kindly have stated that it is a more intuitive and user friendly platform where end users can easily train, test, validate and deploy chatbots. The research and development in this project further this competitive advantage, and will allow Kindly to further market itself as a cutting edge platform that allows companies to build highly complex chatbots in a user-friendly and intuitive manner.

In 2017 Convertelligence secured investments from the Norwegian private investor Erik Must. His long-term investment strategy has ensured a flow of capital to Convertelligence that have allowed the company to scale its operation and grow its sales. Convertelligence will also consider to raise more capital in order to have the option to scale operations at an even larger scale when a more global approach is taken in 2020. The risks involved in scaling operations with external capital is that the return on investments are not met, so outside investments are carefully aligned with Kindly's product-market fit and sales figures. The research and development of this project, if successful, will result in a more mature product with a greater possibility for expanding to markets outside of Scandinavia, at which point several investment strategies can be examined.

As specified in the work packages, the new Kindly 2.0 functionalities will be developed with relatively short,

iterative cycles of development and evaluation. The company partners (Kommuneforlaget, Norwegian Airlines ASA and Adecco Norway) will gain early access to prototype versions of these functionalities through a dedicated test environment. These partners will then provide feedback based on their usage of Kindly, and steer the research and development accordingly. This short feedback loop between the R&D team and the chatbot creators is essential to ensure the technological solutions developed during the project provide value to the end users of Kindly.

For the company partners themselves, the value creation revolves around the possibility of gaining early access to new chatbot development functionalities, thereby allowing them to quickly scale up and enhance their chatbots and provide them with a clear competitive advantage over their competitors.

11. Socio-economic benefits and contribution to sustainable development in society

The empowerment of non-technical users to work with advanced AI technologies stands out as a major contribution of the proposed project for the society at large. AI and ML have become near-ubiquitously used in the daily life of many people, and the participants in this project acknowledge their responsibility to make the technologies they develop as transparent as possible to non-technical users. In addition, by sharing the findings of the research activities (via publications and seminars, see §12), this project will contribute to furthering NLP solutions for Scandinavian languages at national and global scales.

12. Dissemination and communication of results

Both the involved researchers from NR and the Convertelligence ML team have a strong track record of internationally published research. The outcomes of this project will be submitted to NLP and AI conferences, and results and data will be made available on public repositories whenever this is feasible. Convertelligence will also host a series of public seminars in order to share the findings of the project with the local machine learning and NLP community. Furthermore, Kindly customers will be invited to participate in training workshops to be introduced to the new functionalities in Kindly 2.0.

PART 4: Other information

13. Ethical perspectives

One important ethical concern for the project regards the privacy of user messages, which may in some cases include personal information such as person names or addresses. The project will take privacy concerns very seriously and adopt a range of privacy-enhancing measures, such as the use of a de-identification tool already integrated in Kindly to automatically remove potentially sensitive information from the message logs. The chatbot data that will be used during the course of the project will always be stored on secure servers with strict access control.[3] As of today, Kindly is processing approximately 700.000 customer messages each month, in addition to having large annotated datasets created by Kindly's customers.

The participants in this project will not get direct access to this data, but with acceptance from a subset of Kindly's customers, some of this data will be made available for R&D purposes. Some of this data may contain personal information, and strict measures will be taken to ensure that this information is only accessible for the duration of the project, and that guidelines will be provided to all participants on both legal and ethical standards that should be followed when dealing with such information. The project will follow

[3] NR has many years of experience running research projects on databases including confidential/sensitive information and has put in place a technical infrastructure for handling such data in a safe manner.

standard ethical guidelines for scientific research and publications as outlined by the National Committee for Research Ethics in Science and Technology.[4]

14. Recruitment of women, gender balance and gender perspectives

The fields of NLP and machine learning lag behind in gender balance, and special efforts will be made to recruit female applicants for future positions connected to the proposed project at Kindly. Kindly is already in close dialogue with the Language Technology Group (LTG) at the University of Oslo in regards to recruiting female students within our field.

References

Bach, Stephen H., Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, et al. 2019. "Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale." In *Proceedings of the 2019 International Conference on Management of Data*, 362–75. SIGMOD '19. New York, NY, USA: ACM.

Frosst, Nicholas, and Geoffrey Hinton. 2017. "Distilling a Neural Network Into a Soft Decision Tree." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1711.09784.

Galitsky, Boris, and Saveli Goldberg. 2019. "Explainable Machine Learning for Chatbots." In *Developing Enterprise Chatbots*, 53–83. Springer.

Gao, Jianfeng, Michel Galley, and Lihong Li. 2019. *Neural Approaches to Conversational AI: Question Answering, Task-Oriented Dialogues and Social Chatbots*. Foundations and Trends(r) in I.

Jain, Sarthak, and Byron C. Wallace. 2019. "Attention Is Not Explanation." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–56. Minneapolis, Minnesota: Association for Computational Linguistics.

Koh, Pang Wei, and Percy Liang. 2017. "Understanding Black-Box Predictions via Influence Functions." In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1885–94. ICML'17. Sydney, NSW, Australia.

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d16-1230.

Mallinar, Neil, Abhishek Shah, Rajendra Ugrani, Ayush Gupta, Manikandan Gurusankar, Tin Kam Ho, Q. Vera Liao, et al. 2019. "Bootstrapping Conversational Agents with Weak Supervision." *Proceedings of the AAAI Conference on Artificial Intelligence*.

Meena, Raveesh, Jose Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. "Automatic Detection of Miscommunication in Spoken Dialogue Systems." In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Rastogi, Abhinav, Raghav Gupta, and Dilek Hakkani-Tur. 2018. "Multi-Task Learning for Joint Language Understanding and Dialogue State Tracking." In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 376–84. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ratner, Alexander, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. "Snorkel: Rapid Training Data Creation with Weak Supervision." *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* 11 (3): 269–82.

Ratner, Alexander, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. 2019. "Training Complex Models with Multi-Task Weak Supervision." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4763–71.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?'" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*.

Tian, Zhiliang, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. "How to Make Context More Useful? An Empirical Study on Context-Aware Neural Conversational Models." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Zion Market Research. 2018. *"Chatbot Market by Type (Support, Skills, and Assistant), by End-User (Healthcare, Retail, BFSI, Travel & Hospitality, E-commerce, Media & entertainment Entertainment, and Others) by Region (North America, Europe, Asia Pacific, Latin America, and Middle East and Africa): Global Industry Perspective, Comprehensive Analysis, and Forecast 2017-2024*". https://www.zionmarketresearch.com/report/chatbot-market

[4] https://www.etikkom.no/forskningsetiske-retningslinjer/naturvitenskap-og-teknologi/