Grant Agreement No.: 823783
Call: H2020-FETPROACT-2018-2020

Topic: H2020-FETPROACT-2018-01
Type of action: RIA

D4.2 Algorithms and Implementation of Diversity Aware Single user incentive design

Revision: v.2.0

Work package	WP 4
Task	Task Number T4.2
Due date	28/02/2022
Submission date	28/02/2022
Deliverable lead	Avi Segal
Version	2.0
Authors	Kobi Gal, Avi Segal (BGU)
Reviewers	Nardine Osman (CSIC)

Abstract	We report on our three main activities in the reporting period which covers the deliverable focus on incentive design for WeNet with diversity

Co-funded by the Horizon 2020
Framework Programme of the European Union

	awareness and in connection with the entire project. Our first activity is the incentive server developed for the project, its main integration points with other components, and its incorporation into Pilot1, the first pilot run by the project. We describe the deployment of the server to generate incentives for the first WeNet pilot, and report on the results, also comparing them to past literature. We follow this by using the data collected from the M26 pilot for developing and offline testing of a multi armed-bandits based algorithm for incentives personalization. Our second activity is the mutual work with WP9 on Transparency in Machine Generated Personalization, which developed best practices and a checklist for system designers and users of such systems. The developed tools can be used by different audiences to assess transparency in such systems and identify gaps for addressing and coverage. We also describe our planned future work in this important area. We finish our report by describing our third activity on developing recommendation based incentives for increasing motivation and improving productivity and engagement for single users. We focus on our work of designing adaptive explanations based on user's characteristics and behaviour in the system. These explanations are diversity aware by using users' behavioural patterns as opposed to their fixed demographic characteristics. Our implementations and offline trials point to the potential of these approaches and to the promise of designing adaptive explanations for these environments. In this research we work with a large-scale volunteer-based crowdsourcing system which serves as a fertile ground for our future planned experimentations in the wild.
Keywords	Incentive Design, Message Based Incentives, Badge Based Incentives, Recommendation Systems, Adaptive Explanations, Transparency in Personalized Systems

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.8	08/06/2021	Version sent to WP4 internal review	Daniel Ben Zaken, Avi Segal (BGU)
V0.9	14/06/2021	Version sent to WeNet internal review	Daniel Ben Zaken, Kobi Gal, Avi Segal (BGU)
V1.0	27/06/2021	Version after WeNet internal review	Daniel Ben Zaken, Kobi Gal, Avi Segal (BGU)
V1.5	07/02/2022	Revised report sent for WeNet internal review	Kobi Gal, Avi Segal (BGU)
V2.0	28/02/2022	Revised report after WeNet internal review	Kobi Gal, Avi Segal (BGU)

Co-funded by the Horizon 2020
Framework Programme of the European Union

Disclaimer

Copyright Notice

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:		**R**
Dissemination Level		
PU	Public, fully open, e.g. web	✔
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to WeNet project and Commission Services	

** R: Document, report (excluding the periodic and final reports)*

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

EXECUTIVE SUMMARY

Deliverable 4.2 reports on the three main efforts for WP4 in the reporting period which covers the design and deployment of intelligent incentives for sustaining engagement of diverse types of participants.

The first effort consists of the incentive server which is a main component of the WeNet architecture. We provide a detailed description of the incentive server, its main integration points with other WeNet components, and its incorporation into Pilot1, the first pilot run by the project. We also describe its deployment in the first pilot, where the incentive server was used to generate two types of incentives: badges and motivational messages. We provide a detailed analysis of the results of the incentives, analyzing the effect of badges and messages on participant activity in the WeNet platform. We observe that participants increase their level of activity as they near the badge and show that this effect varies across the different communities of WeNet (country and work intensity). We provide an algorithm and simulation for optimizing the incentive design using shallow and deep features of users' interactions that was conducted on data collected from the pilot.

The second effort describes our mutual work with WP9 on Transparency in Machine Generated Personalization. We generated a list of transparency best practices for machine generated personalization. We further develop a checklist to be used by designers to evaluate and increase the transparency of their algorithmic systems. We appllied the checklist to prominent online services and discuss its advantages and shortcomings.

The final effort describes our additional work on incentive design by using explanations to augment recommendations of activities for users. We generate post-hoc explanations for each recommendation which emphasize projects that relate to users' preferences, as determined by their past interactions, using machine learning. The explanations are adapted to the diversity of users by their past behavior and interactions. We demonstrate this approach in an online study on the SciStarter citizen science platform. We provide results that show that users who receive explanations exhibit higher engagement and satisfaction than users who did not receive such explanations. This demonstrates the role of explanations as diverse aware incentives for guiding the behavior of participants.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table of Contents

Co-funded by the Horizon 2020
Framework Programme of the European Union

1. Diversity Aware Incentive Server

The incentive server is concerned with generating diversity-aware incentives to guide people's interactions in the WeNet platform, and is a main component of the WeNet architecture. For a full description of the incentive server API, we refer to WP6's deliverable, as is the case for all technological components of WeNet. In this section we provide a general description of this contribution, covering several topics below.

1. The main functionalities supported by the Incentive Server.
2. The interface points of the incentive server with the other components of the project
3. The incorporation of the Incentive Server into Pilot 1, including the experimentation done, the various conditions tested, and a summary of the main results.
4. A development and testing of a machine learning based algorithm for incentive adaptation based on the (limited amount of) data collected during pilot 1.

1.1 Incentive Server: Main Components and Functionality

The Incentive Server is a standalone server responsible for allowing WeNet applications to guide the behavior of WeNet users and WeNet communities using diversity-aware non-monetary incentives. These applications can be developed by WeNet internal developers (as in Pilot 1, Pilot 2 and beyond) as well as by WeNet's external developers, like the organisations that are going to develop apps on top of WeNet as part of WeNet's Open Call initiative. In fact, at the time of writing this updated report (January, 2022), the Incentive Server is already supporting the Open Call initiative.

Table 1 specifies the main components which are part of the Incentive Server platform:

Table 1: Incentive Server Main Components

Component Name	Component Description
Incentive server IFS	A Django (Python) based web architecture. Manages the tiered application of the server from request reception to request handling to response production. Includes the administration logic of the Incentive Server.
IS Badgr-server	An open-source Python/Django backend for defining and issuing Open Badges (the world's leading format for digital badges). WP4 has adapted and incorporated this open

Co-funded by the Horizon 2020 Framework Programme of the European Union

	source platform for usage as part of WeNet's Incentive Server Solution.
IS Schedler	A Cron based scheduler for managing scheduled tasks for the different incentives and algorithms (e.g. timing of incentive messaging, cleanup tasks and the likes).
IS Administration Application	A Web application for defining and managing Badge based incentives and message based incentives for different WeNet communities and apps. These administration capabilities are also supported through the Incentive Server APIs.
IS User Signals Receiver	Receiving and accumulating information supplied by other WeNet components about users and communities in the system for incentive generation purposes. This includes behavioural and demographic user data containing both shallow and deep diversity features for each user.
IS Rule Engine	Enabling definition of internal rules and norms for granting different types of incentives, as well as restricting such grating by different constraints (e.g. amounts). Such rules can use all available user signals received by the Incentive Server from other WeNet components, including behavioural and demographic signals.
IS Algorithm Interface	Interface for an informed software component (e..g machine learning based) for allocating incentives based on prior developed model and real time user signals
IS web engine	Nginx based web engine for serving incentives through REST based API
IS DataBases	Based on the open source MySQL DataBase. Saving incentive definition information as well as data needed for issuing and tracking incentives.

The current implementation of the incentive-server supports two types of incentives: Badges and motivational messages. We expand on each of these mechanisms below. In section 1.3 we will describe the specific incentives defined for Pilot 1.

Badge Incentives:
This incentive type uses streamed data from other WeNet components and analyses it towards issuing predefined badges to WeNet users. Badge issuing is incentive aware in

Co-funded by the Horizon 2020 Framework Programme of the European Union

that the badge type can be personalized to different types of users. The component supports the following services:

- Badge creation, modification, and removal: Allowing the execution of basic operations on badges using the badgr-server open source code
- Badge allocation through the Incentive Server REST API
- Badge delivery: Allowing the WeNet applications to access badge graphics and badge textual description, presenting them to the end user through various mediums.
- Badge types and badge allocation rules are defined a-priori through the Incentive Server admin application or REST APIs.
- Badge behaviour logging and badge allocation logging.

Incentive Messages:

This incentive type generates motivational messages presented to users by WeNet applications. The messages can be adapted to fit diversity in both behavioural and demographic aspects of users. User data is received and accumulated by the Incentive Server and used by this component. Different user behavioural characteristics are computed such as:

- activity patterns per task and per multiple tasks
- time spent on task
- inactivity periods
- behavioural changes in activity patterns across time

These characteristics are used to guide and personalize the message based incentives. The component supports:

- Incentive messages creation, modification, and removal
- User behavioural logging and message sending logging
- Synchronization with badge allocation
- Incentive messages notification: Notifying the users about incentive messages while taking into account the user's overhead pre-defined conditions (e.g not allowing multiple incentive messages from the same type in a specific time window)

1.2 Main Integration Points

We now describe the main integration points with other WeNet components used by the Incentive Server during Pilot 1.

Co-funded by the Horizon 2020
Framework Programme of the European Union

TABLE 2: INCENTIVE SERVER MAIN INTEGRATION POINTS

Component	Method	Data	Uses
WP5 - interaction protocol	POST	Incentive Information	The Incentive server informs the users about a new incentive issued. This is done through the interaction protocol which also verifies the user's ability to get notifications. (i.e. there may exist a norm in the interaction protocol component delaying such interaction)
WP5 - profile manager	GET	User Metadata	The incentive server uses the user's profile metadata as part of its data for generating incentive messages. For example, based on time in the system, preferences on time of day for receiving messages etc. Such user information is supplied by the profile manager.
WP5, WP2, WP3 - tasks manager and additional information	GET	Users' tasks and TasksTransactions information	The Incentive Server is querying the task manager for data regarding users behaviour in the system. These data points are being analyzed to calculate the user's criteria towards an incentive. Behavioral information is supplied to the task manager also from WP2 (single user information) and WP3 (social interactions - future). For example, if the user joined

Co-funded by the Horizon 2020
Framework Programme of the European Union

			1 week ago and did not ask any questions, the incentive server may issue an incentive to encourage the user to ask a question.
WP6 - platform	Multiple	Incentive information Component behaviour information	The Incentive Server supplies the platform with all the user's incentive badges, badges for specific apps and all available incentive messages. Additionally the Incentive Server receives general services from the Platform, such as presenting Incentive Information on the WeNet Hub Page, Logging of overall component behaviour, and access to Incentive Server general behaviour data from an external interface.

1.3 The Incentive Server @ Pilot 1

We now move to describe the use of the Incentive Server as part of Pilot 1 of the WeNet platform. In this pilot users were able to ask questions, give replies to questions and indicate their acceptance (or rejection) of answers supplied to them. Pilot 1 was used by WP4 to test the Incentive Server in the wild, compare between different incentive types, check the influence of various diversity aspects on users' response to incentives and for the collection of data for future algorithmic development.
Table 3 presents the different locations of the pilots and the dates the pilot was run at each location:

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 3: Pilot 1 Locations and Dates Avalilable for Data Analysis

Location	Dates
Denmark	12-28 March 2021
UK	12-28 March 2021
Mongolia	12-28 March 2021
Paraguay	17-31 March 2021

We note that the data from the Italian pilot was not available at the time of this analysis and thus not included in the analysis that follows.
In each of the locations above, the following incentives were supported:

Badge Incentives (each cell in the table also shows the badge graphics as presented to users) :

Table 4: Badge Types in Pilot 1

Badges for Asking Questions	**Badges for Answering Questions**	**Badged for Having one's Answer Accepted**
First question asked: badge granted after user asked a first question	First answer given: badge granted after user answered first question	First answer accepted: badge granted after user's first answer was accepted

Co-funded by the Horizon 2020 Framework Programme of the European Union

		
Asked 5 questions: badge granted after user asked 5 questions 	Answered 5 questions: badge granted after user answered 5 questions 	3 answers accepted: badge granted after user's 3rd answer was accepted 
Asked 10 questions: badge granted after user asked 10 questions	Answered 10 questions: badge granted after user answered 10 questions	5 answers accepted: badge granted after user's 5th answer was accepted

Co-funded by the Horizon 2020
Framework Programme of the European Union

The badge types and thresholds were defined in consultation with WP7.

Additionally, in each of the locations, the following incentive messages were supported:

Incentive Messages

Table 5: Incentive Message Types in Pilot 1

Badge steering Messages	Personal Focus Messages
“You are x questions away from a new badge!”	“You haven't asked a question yet. You can get help from the community with your questions.”
“You are x answers away from a new badge!”	“There are open questions to answer. Type /answer for the list!”

Co-funded by the Horizon 2020
Framework Programme of the European Union

Where “x” is computationally added by the incentive server upon deciding to serve a message.

The badge steering messages were meant to be used in conjunction with the badge incentive and measure the impact of using both mechanisms in cooperation. The personal focus messages were used to remind users about the main actions available in the system. Both messages were served at semi-random intervals under the following norm limitations in order to collect data for developing an informed algorithm (described in section 1.5):

- send x times: Maximum 6 times during pilot
- frequency: once in random(1, 4) days
- After how much inactivity: 1 days from join/ inactivity.

These norms were set in consultation with WP7 and WP5.

Collection of data with semi-random policies (i.e. manipulating the message sending frequency at random) enables future development of machine learning based algorithms that can maximise the incentive trajectories over time. This is achieved since these semi-random interventions constitute an exploration phase where the various incentives effectiveness is measured in different system states [1]. Using norms to limit this randomness ensures that we protect users and do not serve incentives entirely out of context.

Research Questions: Effect of Incentives

Past research has demonstrated that both the timing of incentive messages as well as their content is critical for incentivising users to increase their contributions in the system[2,3]. In this specific experiment in the wild, we seek to check the following research questions:

1. What is the effect of badges on participant activity in the WeNet platform. In particular, do badges generate a ‘’steering’’ effect by which participants increase their level of activity as they near the badge [4, 5]?
2. Is the effect created by badges different for different communities and groups as organised by shallow diversity features (e.g. group per location) or deep diversity features (e.g. group per level of activity)?
3. Is there a decline in contribution following the steering effect as identified by prior research, and is this decline also different for different groups [5]?
4. Do message incentives on top of badge incentives further increase contribution by users?
5. What are the opinions of users about the incentives used, badges and messages, following the experiment?

Additionally, this experiment was used as a data collection step to enable future incentive policy development in followup research steps.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Towards these goals users were allocated to two cohorts incentive wise at each one of the above locations. The cohorts included: (a) Cohort 0: users allocated to this cohort received only badges as incentives in the system (b) Cohort 1: users allocated to this cohort received badges as well as incentive messages in the system.
We note that the relatively small expected amount of users per location prevented a definition of a 3rd cohort where no incentives are used.

The data collected during the experiment included time stamped data about users' activity (questions asked, answers given, answers accepted), the incentives used, as well as survey and interview data following the experiment.

We now describe the main finding of our experiment during Pilot 1.

1.4 Analysis of Pilot 1 Results

Figure 1 details the number of participants that participated in each pilot (where data was available for analysis at the time of this report writing):

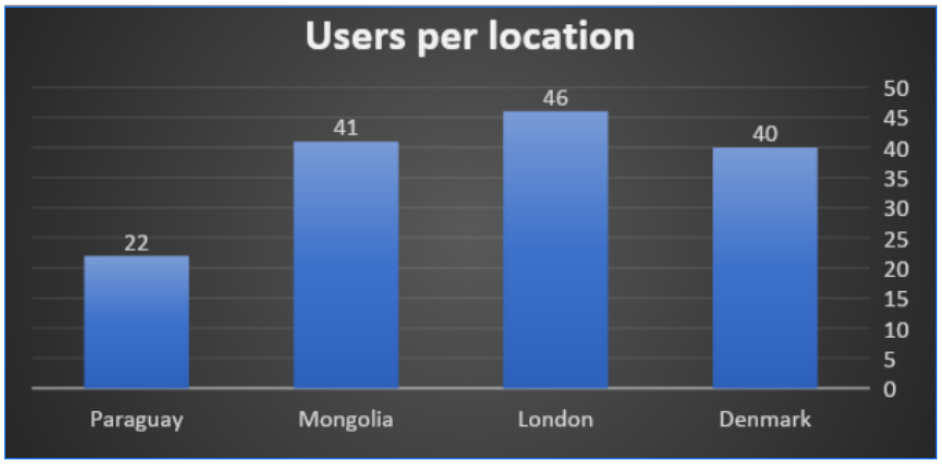

Figure 1: Users per pilot location

We note that the UK pilot (London) had the largest number of participants (46) while the Paraguay pilot had the lowest number of participants (22).

Figure 2 presents the total number of badges allocated and incentive message used across all locations, per cohort:

Co-funded by the Horizon 2020
Framework Programme of the European Union

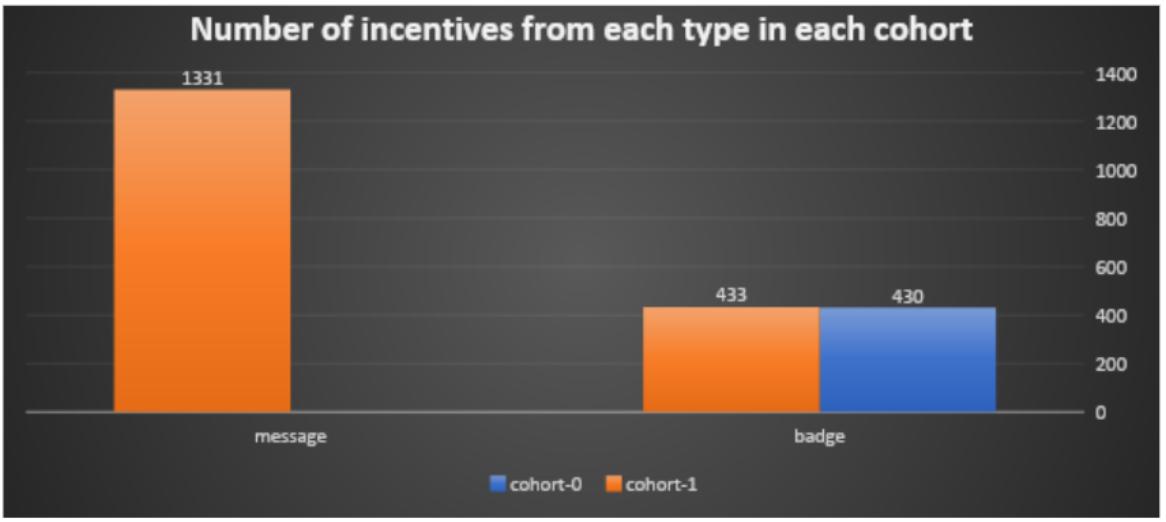

Figure 2: Incentives per type and cohort

Per the figure, in cohort-0 only badges were used, and overall 430 badges were allocated, while in cohort-1, 433 badges were allocated and 1331 incentive messages were sent.

To answer research question 1, we check for users' behaviour prior to receiving a badge and following a badge reception. We first look at the average behaviour across locations and users actions. As can be seen in Figure 3 below, steering is indeed identified in users' behaviour. Users are increasing their actions towards the badge allocation day and decreasing their actions following the badge granting. Interestingly, the decrease is to a new level of activity, higher than the one prior to the badge grating.

Figure 3: Badges steering effect - overall

Co-funded by the Horizon 2020
Framework Programme of the European Union

To check research question 2, we move to investigate the steering per location, as location constitutes a shallow feature of diversity. We notice in Figure 4 below a unique steering behaviour per location. Specifically, users in the UK demonstrate the smallest steering effect, while users in Paraguay are strongly steered by badges, but then demonstrate a sharp decline. This finding of different responses to badges per location indicates the need to consider each location in separating for badge design in diversity aware systems.

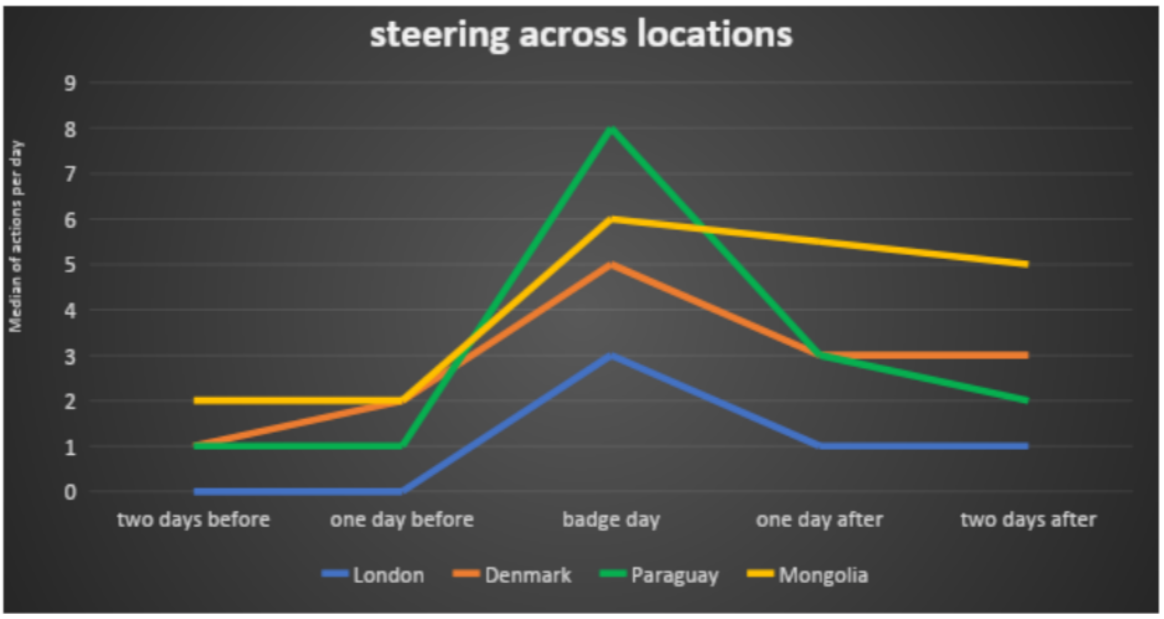

Figure 4: Badges steering effect per location

To further investigate steering across diversity dimensions, we also check the steering behaviour for a deep diversity feature which is different from users' demographics. For this, we divide users in the systems to two types: (1) low - low intensity working users with 4 or less actions in the systems prior to receiving a badge (2) high - high intensity working users with more than 4 actions in the system before receiving a badge. Working intensity constitutes a deep diversity feature which relies on the user's capabilities as opposed to their demographic features. Figure 5 presents the steering behaviour of the two groups. As can be seen in the figure, high intensity working users are much more influenced by badges on average than low intensity working users. This indicates the need to take such deep diversity features into account when designing incentives, e,g by augmenting the incentive for low intensity working users with additional motivational factors.

Regarding research question 3: We found that contribution does decline following the steering effect (as identified by prior research), and this decline is specific per group. This

Co-funded by the Horizon 2020
Framework Programme of the European Union

is the case for groups separated by location (shallow diversity features) as well as work intensity (deep diversity feature).

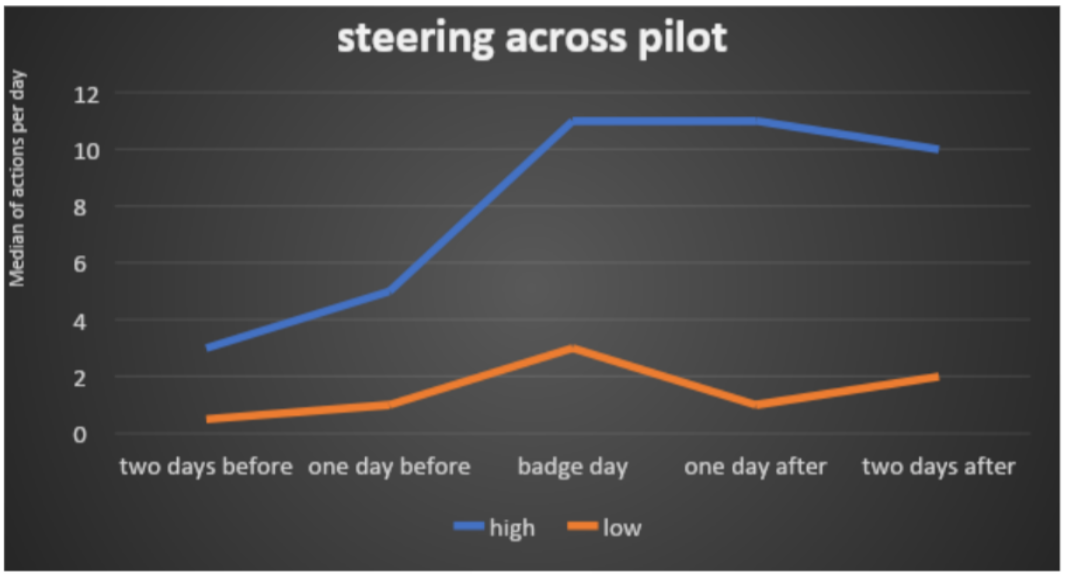

Figure 5: Badges steering effect per working intensity

We now move to investigate research question 4, checking whether message incentives on top of badge incentives further increase contribution by users.
Table 6 outlines the total actions performed by users in each of the locations, when actions include the 3 operations possible by a user: asking a question, answering a question or accepting an answer.

Table 6: Total Actions Per Location and Cohort

Location	Cohort-0: Only Badges Total Actions	Cohort 1: Badges + Messages Total Actions
UK (London)	554	**653**
Paraguay	453	**842**
Denmark	952	**1235**
Mongolia	**2437**	1918

Co-funded by the Horizon 2020
Framework Programme of the European Union

As seen from the table, in three of the four locations - the UK, Denmark and Paraguay, indeed the cohort receiving both incentive types demonstrated higher productivity. Interestingly, the observation in the Mongolian pilot is the opposite, with the badge only group demonstrating higher responsiveness and productivity. This research question should be further investigated. We hypothesise that a diversity factor may have impacted these results, be it a shallow one (such as location) or a deeper one, possibly connected to users' activity level (as can be seen from the table, the mongolian pilot was the most productive one across cohorts).

Finally, we checked users' opinions about badge and message incentives, as demonstrated in the post experiment exit surveys and interviews.

The exit surveys administered at each location included the following questions regarding incentives:

1. "Please write in at least one badge you received." (to verify if users remember this incentive type)
2. "Please indicate whether you agree or disagree with these statements" (Likert scale of 1-5 was used, with "1" indication strong disagreement and "5" indicating strong agreement)
    a. "I liked the chatbot's badges"
    b. "The badges were a distraction"
    c. "The badges enhanced the chatbot experience"
    d. "The badges encouraged me to contribute to the chatbot"
    e. "Chatbot should be more generous with badges"
    f. "More type of badges should be used"
    g. "Badges based on the acceptance of answers should be used more"
3. Please write in at least one message you received (Question presented only to cohort 1 users. Used to verify if users remember this incentive type)
4. "Please indicate whether you agree or disagree with these statements" (Questions presented only to Cohort 1 users. Likert scale of 1-5 was used, with "1" indication strong disagreement and "5" indicating strong agreement)
    a. "I liked the chatbot's messages"
    b. "The messages enhanced the chatbot experience"
    c. "The messages were a distraction"
    d. "The messages encouraged me to contribute to chatbot"
    e. "More types of messages should be used"
    f. "Messages should be sent less frequently"
    g. "Messages should be personalised for each user"

Figure 6 presents the results of questions 1 and 3 above. Specifically, we count for each incentive type and each location the percentage of users who were able to remember at least one incentive which they have received.
We note that due to a collection error, the message incentive information from the Paraguay location is not available. Thus, for the message incentive type, only data from the UK, Denmark and Mongolia are presented.

Co-funded by the Horizon 2020
Framework Programme of the European Union

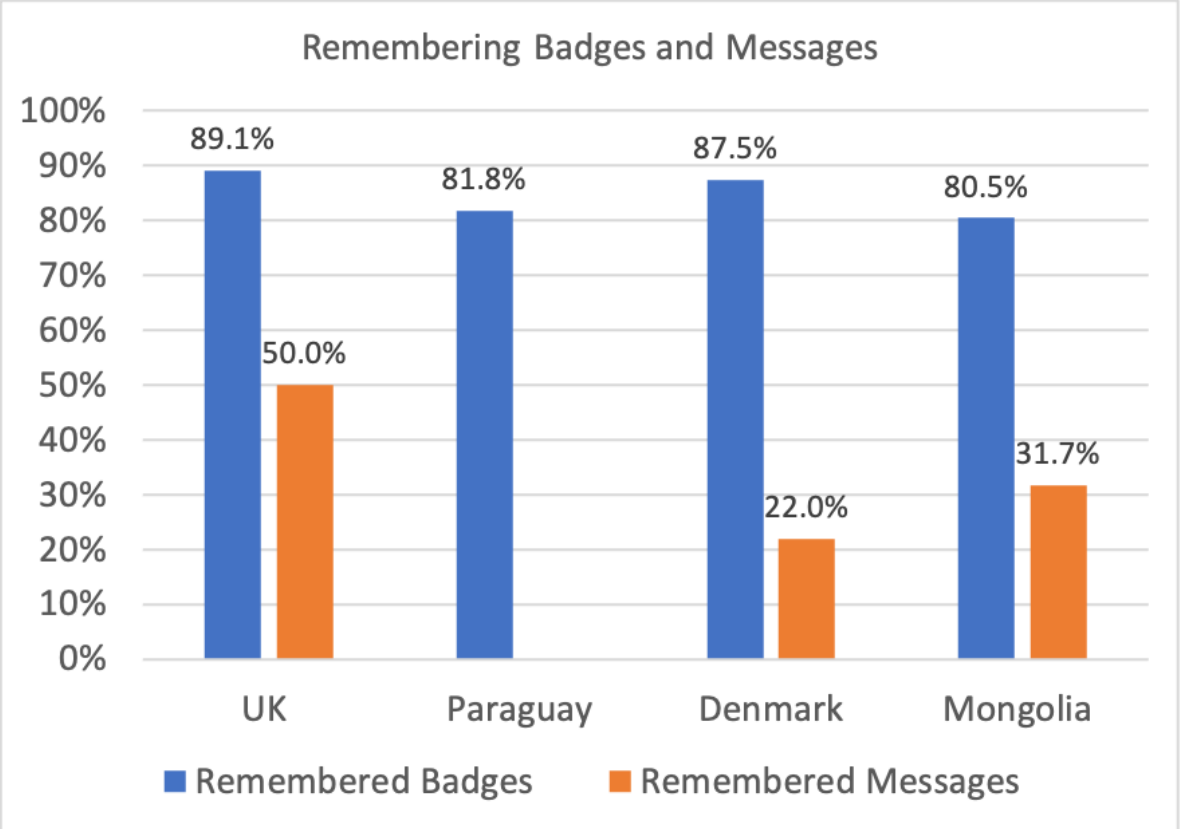

FIGURE 6: INCENTIVE REMEMBERING ACROSS LOCATIONS

We note the following insights from this analysis: (1) users were able to remember badge incentives much more than message incentives across locations. In fact, at least 80% of the users across locations remembered at least one badge they have received, while only 50% of users or lower remembered an incentive message that they have received. We hypothesise that one factor which contributed to this phenomena was the fact that the system sent many additional messages to users across its operation (besides incentive messages), thus rendering the incentive message less memorable. We conclude that incentive messages should be better separated from other system messages and that possibly, the frequency of other system messages should be decreased. (2) We note the differences between the different locations in remembering badge incentives and message incentives. For example, only 80.5% of users remembered at least one badge in Mongolia while 89.1% of users remembered such badges in the UK. This indicated that auxiliary mechanisms of attention should possibly be used to account for such differences between diverse populations (e.g. by phrasing the badge language differently, communicating differently the badge allocation, etc.).

Co-funded by the Horizon 2020
Framework Programme of the European Union

Figure 7 and 8 present the average ratings of users to the badge and messages rating questions across locations.

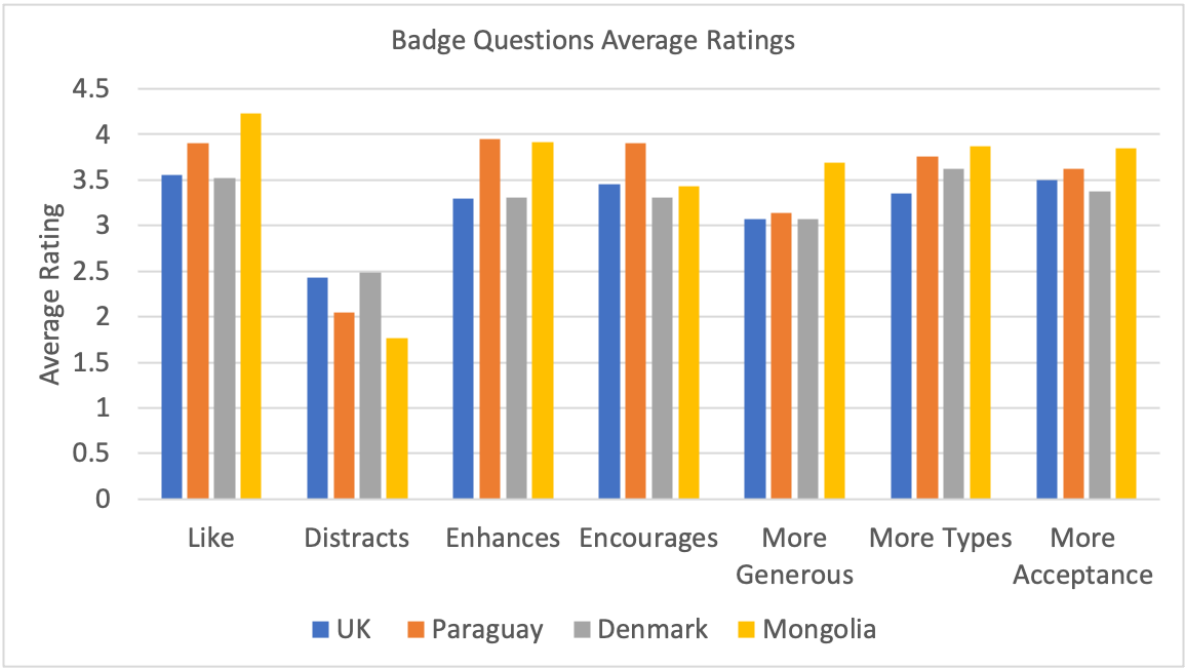

FIGURE 7: BADGE OPINIONS ACROSS LOCATIONS

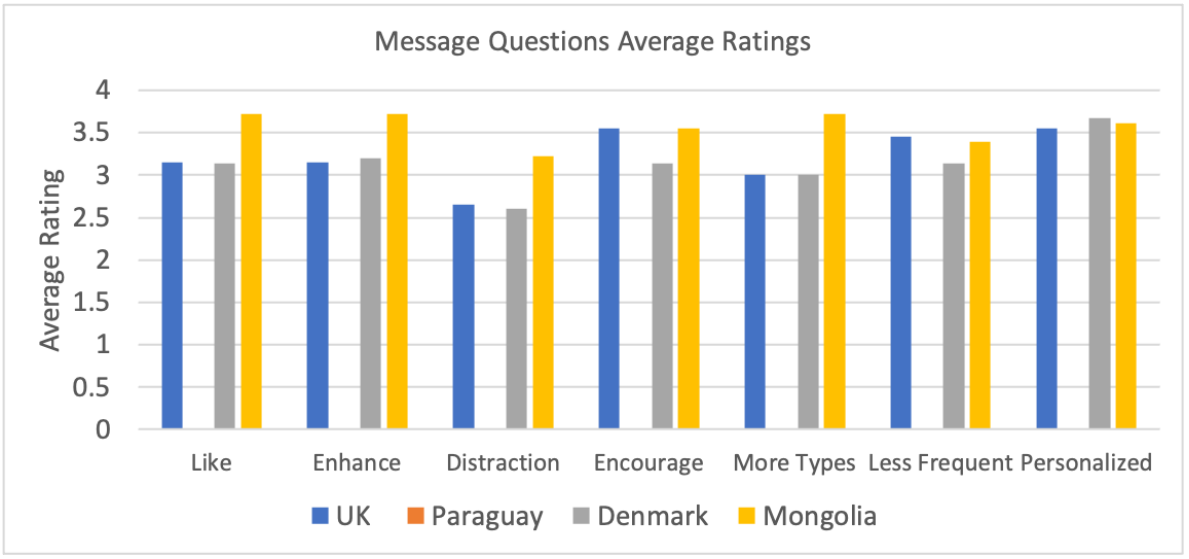

FIGURE 8: MESSAGE OPINIONS ACROSS LOCATIONS

Co-funded by the Horizon 2020
Framework Programme of the European Union

We can observe several trends from the two graphs: (1) On average, users expressed more liking toward badges compared to messages and found them less distracting. (2) This preference is evident in other responses as well (in the graph), i.e. regarding how much badges enhance performance, encourage contribution, and whether users wanted to see more types of the incentive types. For all these questions users give higher ratings to badges compared to messages. (3) Users' average opinions are location dependent. For example, users in Mongolia expressed stronger liking to badges then users in other locations and this liking is also evident in their other replies. We note the relatively high ratings users in the UK gave to badge and message distraction questions, which is in line with their opinions as expressed in the focus groups. (4) Overall, users expect more badge types and more quality based badges (e.g. like the “your question has been accepted by another user” badge). (5) Overall, users expect messages to be less frequent, and would like to see messages personalised for them.

In the exit focus groups, users were asked for their opinions about badges and messages in the application.

Users in the UK, Denmark and Paraguay had the opinion that badges had some motivational effect but said that there was not enough information about the badge granting mechanism. We hypothesise that this is a result of having information about the badge types and thresholds only on the WeNet hub and not in the app itself. Users in these locations also have had additional ideas for mechanisms that will improve motivation such as adding additional awards for obtaining a badge (UK), having a weekly challenge (Denmark) or adding competitions between users (Paraguay). Additionally, users in the UK and Denmark felt that badges should be better tied to and presented at the user profile to showcase the users’ achievements. Interestingly, a UK user mentioned the risk of discrmination if some users with less badges on their profile will be treated differently than more active users.

Users in Mongolia found badges very motivating, and mentioned that they felt that receiving a badge was like winning a competition. Some users also mentioned that the offered badges were not challenging enough and that they expected more badges as they continued to contribute. This is well represented in the data, as users in this location outperformed all other locations in terms of actions undertaken in the app, “exhausting” their badge inventory early on. Users in this location also found messages the most distracting, compared to other locations. The strong liking of badges in this group combined with the high level of perceived message distraction, may explain the fact that messages did not improve the performance of users in this location. We hypothesise that cohort-0 users in this location, which received only badges, were not distracted by messages, and continued to contribute expecting more badges further along the way.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Taken together, these results demonstrate that different user groups have different expectations from incentives, hold different opinions about how incentive mechanisms should behave and about what is the influence of such mechanisms on them. Additionally, as demonstrated by the steering results, different user groups are steered differently by incentives, and respond differently to the combination of incentive mechanisms (in our case badges and messages). We showed that at least for badges, this difference in steering holds both for groups defined by a shallow diversity feature (location) as well as groups defined by a deep diversity feature (activity level). These findings are in line with the need to further develop personalised incentive mechanisms which adapt per user's diversity features, be it shallow or deep.

Co-funded by the Horizon 2020
Framework Programme of the European Union

1.5 Diversity-Aware Incentives Based on Machine Learning

In this section we use the collected data from the M26 pilot to develop a diversity aware machine learning based personalization approach for incentive messages. Specifically, we compare between a non personalised approach for incentive design, and 3 personalization based approaches: one based on the shallow diversity notion, a second based on the deep diversity notion and a third based on a combination of the two.

1.5.1 Optimising Engagement

We focus on the problem of optimising engagement in the system by presenting one of the incentive messages used in pilot 1 to the users. Given the short duration of the pilots (2 weeks), we ask the following question: given the activity of the user during the first week of the pilot, what message should be presented to them at the beginning of the second week (if at all) to maximise their contribution in the second (and last week) of the pilot. This question is rendered important as the analysis done by WP7 demonstrated that activity in the 2nd week of the pilot decreased across locations.

1.5.2 Shallow and Deep Diversity Features

We choose the following two features as information available for the optimization algorithm about each and every user in the system:

1. Shallow feature - location: the algorithm is aware of this demographic information about each user in the system (one of the 4 locations available in the dataset).

2. Deep feature - activity in week 1: the activity of the user in week 1 (questions asked, questions answered, questions accepted) is a deep diversity non-demographic feature about the user which is available for the optimization algorithm.

1.5.3 A Multi-Armed Bandits Approach

We use a multi-armed bandit approach [1] for our incentive messages decision making problem. This family of models is specifically adequate for sequential decision making problems where one wishes to maximize the cumulative sum of **rewards** received over some time horizon. In such environments, the algorithm must choose between several

Co-funded by the Horizon 2020 Framework Programme of the European Union

alternatives (referred to as **actions** or **arms**) and needs to balance between exploiting the information it already has about the actions taken, and exploring additional actions to learn more about the environment. The traditional algorithms in this domain consider the actions selected and the outcome generated from each such selection as input for their optimization processes. We are the first to use multi-armed bandits for the the incentive design task.

Contextual multi-armed bandits [6,7] are a family of algorithms where the **context** of the environments is also taken into consideration (in addition the the action chosen, and the reward obtained from taking this action). Such context can be useful for decision making e.g by identifying different characteristics of users (in the user profile) that influence their response to specific actions, etc. In this setting, the algorithm needs to decide at each time step which arm (action) to select given the context available to it. See Figure 9 below for a schematic diagram of this process.

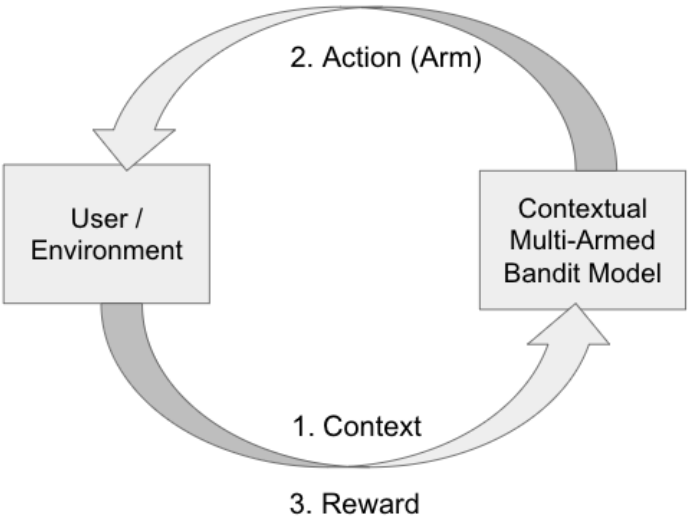

FIGURE 9: CONTEXTUAL MULTI-ARMED BANDIT

Contextual bandits have recently been extensively used to deal with personalization challenges across domains . This included among others personalising web pages, personalising news to users according to their interests, personalising ads to users across the web and adapting learning materials to students [8,9]. We extend such approaches to the socially focused domain of the WeNet project.

Under this framework, we define the incentive optimization problem on the M26 pilt data as follopws:

a. **Learn an optimised policy (i.e. what action to take for each user given a context) on a subset of this dataset, for the following setup:**
    i. **Actions** (arms) - one of the messages used in M26 pilot or no message at all
    ii. **Reward** - level of week 2 activity at the end of week 2. For simplicity we define a binary reward based on the median number of actions performed by users at week 2. We seek to generate contributions of at least 25% over this median threshold for week 2. Under this definition, the reward is 1 if the user performs 10 or more activities in week 2 (the median of week 2 actions is 8) and 0 otherwise
    iii. **Context**: we compare between 4 cases of contexts:
        1. **No context:** the algorithm optimises messages for the beginning of week 2 only based on the historical information of actions taken and rewards obtained
        2. **Shallow**: the algorithm uses the location information in addition to reward and action information
        3. **Deep**: the algorithm uses only the week 1 activity information in addition to reward and action information
        4. **Combined**: the algorithm uses both the shallow and deep information available in addition to the action and reward information

b. **Evaluate this new policy in an offline setting on another subset of the available dataset**

We note the need for offline evaluation (i.e. testing the optimized policy on a test set) when access to online experimentation (i.e. testing the policy in the field) is costly or unavailable as in our case. Even when access to such experimentation systems is available, such offline analysis should be performed as a first step for human centric systems, so as to root out any risky and inefficient action policies before experimentation in the wild.

For the offline evaluation we use the doubly-robust estimator [10] which has demonstrated SOTA results for offline estimation in multi-armed bandits and reinforcement learning based settings.

1.5.4 Evaluation and Results

Our contextual bandits implementation is based on the offset tree algorithm [11] which is one of the recommended methods of choice for learning in offline settings (as in our case). We set the training dataset at 75% of the available M26 data, and the testing dataset (used for the offline evaluation) at the remaining 25% of the data. In our tests each condition was run in separation, manipulating only the context available for that condition as part of each test.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Figure 10 presents the results of the offline evaluation. The X axis presents the different conditions, and the Y axis presents the average accumulated reward multiplied by 100 (to represent percentages). The Y value can be interpreted as the percentage of cases where the algorithm was able to intervene with an action that encouraged the user to perform 10 or more activities in week 2.

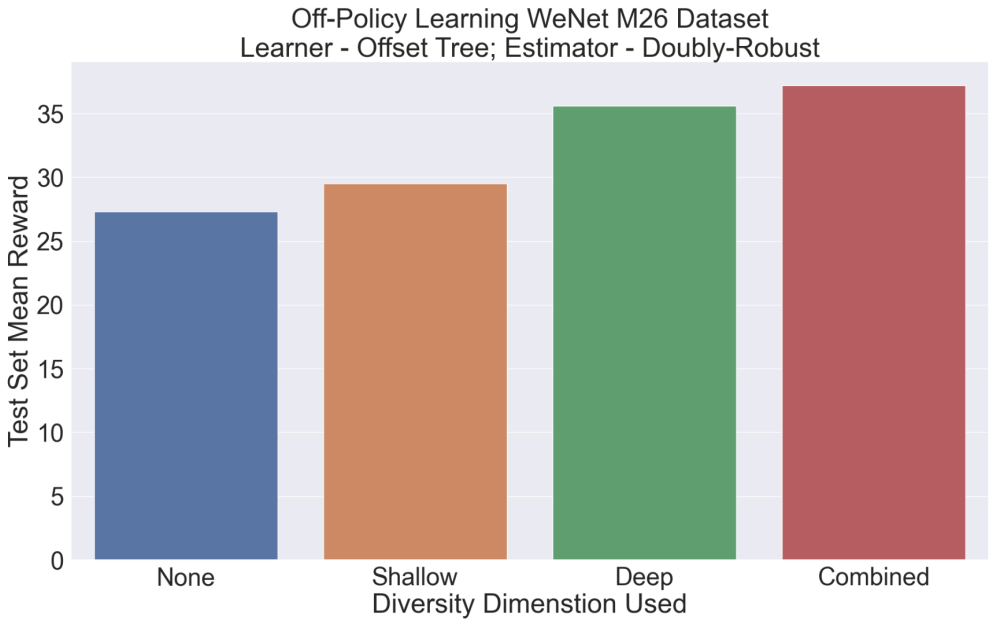

FIGURE 10: OFFLINE POLICY EVALUATION – DIFFERENT CONTEXTS

As can be seen in the figure, the lowest performing algorithm was the algorithm which received no context. This algorithm was able to influence users in 27.3% of the cases. We note that the algorithm which used the deep diversity feature outperformed the shallow diversity based algorithm, achieving 35.6% influence, compared to 29.5% influence achieved by the algorithm using the shallow information. Finally, the algorithm combining the two features was able to outperform all other alternatives, demonstrating an influence of 37.2% of user cases.

These results, while preliminary and based on the small dataset available so far, shed a light on the potential of personalising incentives based on a varied set of user diversity features, deep as well as shallow. They also hint as to the benefit of using deep diversity features for such personalisation, which have demonstrated the potential of improving outcomes on top of shallow features.

Future work based on additional collected data (in Pilot 2 etc.) will seek to optimise additional reward metrics (beyond 2nd week activity) and to use more available shallow

Co-funded by the Horizon 2020
Framework Programme of the European Union

and deep diversity features as context. Additionally, we will seek to test our resulting optimized policies in the wild, beyond the offline settings demonstrated above.

Co-funded by the Horizon 2020
Framework Programme of the European Union

2. Best Practices For Transparency In Machine Generated Personalization

In this section we briefly report on our joint work with WP9 on developing best practices and a checklist for measuring and implementing transparency in machine generated personalization. We note that this work culminated in a joint publication in the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020 (the link for the paper is here, and an introductory video on our joint work is available on the bottom of that page). We also briefly describe our next steps following that publication.

The motivation for this joint work stemmed from the observation that machine generated personalization is increasingly used in online systems. System designers, data scientists and developers are using machine enabled personalization to provide users with relevant content, products, and solutions that target their respective needs and preferences. Such approaches are also used to maximise systems goals, i.e. by focusing on increasing users engagement, involvement, and spendings. However, the same approaches might leave users vulnerable to online manipulation due to algorithmic advancements and lack of transparency. Such manipulations may decrease users' levels of trust, autonomy, and satisfaction concerning the systems with which they interact.

Thus, WP4 and WP9 have worked together to identify the main aspects of transparency which are of importance ethical wise and implementation wise in these personalization based systems. We specifically focused on developing tools to guide system designers implementing transparency in their systems.

In our joint work we have combined insights from technology ethics and computer science to generate a list of transparency best practices for machine generated personalization. We have further developed a checklist to be used by designers to evaluate and increase the transparency of their algorithmic systems. Adopting a designer perspective, we applied the checklist to prominent online services and discussed its advantages and shortcomings.

As stated above, the details of this work can be reviewed in our joint publication.

We describe here our additional work following that publication. In this followup work, we have built a large scale study for collecting users' expectations as to the transparency features and levels they would like to have in personalised systems, and their opinions as to actual levels of transparency they encounter today in online systems. The link for this online study can be found here. In this joint work we have focused on the experiences and expectations of users about transparency in five commonly used online systems that use AI to generate personalized recommendations. We surveyed 108 students who use services provided by these systems, asking them (1) to assess the level of transparency that is currently exhibited by these services; (2) to provide their own preferences and opinions about different aspects of transparency in online systems. Our data collection has now finished and the results of this study will be submitted for publications jointly by WP4 and WP9 and will be reported in the next reporting period.

Co-funded by the Horizon 2020
Framework Programme of the European Union

3. Generating Recommendations with Post-Hoc Explanations for Citizen Science

In our past report (D4.1) we have developed a new approach for personalized recommendations in citizen science platforms, in order to increase their engagement and motivation to contribute. Our recommendation system delivered personalized recommendations to signed-in users by recommending them with new projects based on their past history on the site and based on projects' characteristics. We applied this approach in the wild in the SciStarter platform (https://scistarter.org). SciStarter offers more than 3,000 projects and recruits volunteers through media and other organizations, bringing citizen science to people. Our results pointed to the potential of the proposed recommendation based approach and algorithms to incentivize users in voluntary non-monetary domains, such as the WeNet system.

In this report we extend our work by using data from our past study to develop explanation algorithms for recommendation systems. Specifically, we develop post hoc explanation approaches for black box recommendation algorithms which are hard to explain. This approach generates adaptive explanations while addressing different characteristics of the user's diverse behaviour in the system. We show that our proposed algorithm outperforms alternative approaches in offline settings and report on our plans of testing these approaches in the wild.

Co-funded by the Horizon 2020
Framework Programme of the European Union

3.1 Related Work

We first mention results showing that providing explanations to users in recommendation systems can improve the system's transparency, persuasiveness, effectiveness, trustworthiness, and user satisfaction [12,13]. McAuley and Leskovec [14] extracted topics from Amazon's products, and used related product reviews to explain the decisions of a latent factor model. This improved product rating prediction accuracy as well as generated explanations to recommended items. Bilgic and Mooney [15] have shown that providing explanations help users discard irrelevant options while allowing them to recognize good ones.

Various forms of explanations for recommendation have been explored in prior work, including sentences, word clouds, as well as different kinds of visualizations [16,17]. A common approach in large scale systems (e.g., Amazon, Netflix) is to visualize recommendation items in groups (or carousels) such that each group is associated with an explanation. Felicioni et al. [18] compared several methodologies for evaluating different carousel designs using offline data.

A significant strand of research has developed models that work post-hoc, in that they receive as input a set of existing recommendations and subsequently generate a justification for each of them [19,20]. This is because state-of-the-art recommendation algorithms use latent factors that are not naturally interpretable by people. In this case the explanations are generated independently and the original recommendation algorithm is a black box. Musto et al. [21] used natural language processing and sentiment analysis techniques to generate post-hoc explanations from users' reviews. Shmaryahu et al. [22] designed methods for generating post hoc explanations given a set of Matrix Factorization based recommendations, which was evaluated on offline data. Their method generates different explanation types by running a set of explainable recommendation algorithms which provide a score for the items recommended by the black box. We extend their work by studying grouped recommendations and by deploying our approach in the real world. Lastly, our work extends previous work on generating project recommendations in citizen science using AI [23] that did not provide explanations to users.

3.2 Methodology

We now move to describe our approach of improving project recommendations in citizen science by combining a Hybrid-BPR based recommendation approach with explanations supplied for each recommended project. Our approach consists of the following steps: (1) We create a Hybrid-BPR recommendation algorithm that combines a content based recommendation approach with the BPR Matrix factorization model [24]. This model relies on latent factors and does not align projects with explanations. (2) We align each of the projects outputted by the Hybrid-BPR algorithm with one of five explanation types (associations rules based, feature based, item-item based, popularity based, and location based). (3) We group the projects based on their derived explanation type and choose an optimal ordering over these groups.

We evaluate the methodology in steps (1)-(3) by conducting off-line experiments as well as a randomized controlled study in SciStarter which compared a system that provided grouped

Co-funded by the Horizon 2020 Framework Programme of the European Union

explanations to projects with a recommendation system without explanations. We now move to describe each of the above steps in detail.

3.2.1 The Hybrid BPR Approach for generating recommendations

We describe our chosen approach for generating recommendations in Citizen Science. It extends a well-known Matrix Factorization approach called BPR (Bayesian Personalized Ranking) algorithm [25,26] with a content-based component. We provide an empirical evaluation of this Algorithm in the next section.

The Hybrid-BPR model combines a collaborative filtering approach using implicit user-projects sparse data with meta information available for the projects in SciStarter. Specifically, this meta information includes the estimated time to complete an activity in a project (in seconds), whether the project's activities are to be carried out indoors or outdoors and the topics associated with each project. These features are then encoded by one-hot encoding and the cosine similarity metric [27] between feature pairs is calculated by measuring the angle between vectors.

The Hybrid-BPR recommendation model generates a list of N ordered recommendations for each user by combining this content recommendation score and the BPR recommendation score for each <user, project> tuple. Specifically, it computes a weighted average of both scores with weights of 0.8 and 0.2 for BPR and content, respectively. These weights were optimized to achieve a maximal precision@3 metric on the available research dataset, with three being the number of recommendations that the users receive on the SciStarter home page.

3.2.2 Explanation Types

The Hybrid-BPR approach uses latent variables and is not amenable to interpretation by people. In this section we provide five independent methods for generating post-hoc explanations to projects.

Topic Association Rules This method uses associations between project topics to produce explanations to users. First, associations between topics in Scistarter are mined based on Scistarter users' past interactions. This mining generates a list of direct associations between pairs of two topics in the system. Then, for each new project recommendation which needs to be explained to the user, the method looks for an exiting association rule which ties the topic of this recommendation with topics of projects the user already contributed to. If such an association rule exists, this explanation type can be used. For example, a recommended project with a topic of Molecular and Cells may be associated with a past project the user has contributed to with a topic of Health and Medicine. This association will then be used as an explanation for recommending the new project. The explanation phrasing will include the text "Because you contributed to projects with Molecular and Cells topics in the past".
To generate the associations between projects, we use an existing algorithm [28] to mine association rules from SciStarter users' past interactions. We consider association with

Co-funded by the Horizon 2020 Framework Programme of the European Union

support higher than 0.01. This value was set empirically to ensure associations are generated even for topics with low support (i.e. in the long-tail of the system), as suggested by [29].

Feature Based This method uses project features to produce explanations [30]. We have access to a set of project features, namely the project description, its topics, estimated time of contribution etc. We then learn the user preferences over these features, and use these computed preferences to generate a top list of new projects with similar features that the user may like. Once we have this list, we check if the project requiring explanation is included in the list. In such a case, an explanation may be generated based on feature similarity between the projects that the user has already contribute to and the newly recommended project for this user. Thus, a possible explanation in this method will be “You were interested in Earth&Life topics in the past". This can be generated in case the user prior contributions and the recommended projects share the same Earth&Life topics.

CF-Item-Item In this method, we use information about the users’ interactions with SciStatrter’s projects to compute similarity between projects [31] based on other users that contributed to the same projects. The cosine similarity metric is used as the similarity metric of choice. The output of this method is a ranked list of the N most similar projects to the user’s prior projects. A similarity score for each such project is also computed. Once this list is available, the method checks if the recommended project requiring explanation is included in it. In such a case, this explanation type may be used. For example, 90% of users who have contributed to the “Instant Wild - Stepnoi Surveillance” project for monitoring Russian Steppe wildlife, have also contributed to the project “Instant Wild: Croatia” for monitoring wolf populations in Croatia. Thus, a user who interacted with “Instant Wild: Stepnoi” in the past and now is recommended with “Instant Wild: Croatia” may receive the following explanation: “People who liked Instant Wild: Stepnoi Surveillance also liked Instant Wild: Croatia.”

Popularity This method uses the popularity of the recommended project in the system as an explanation, in case this project is indeed popular in SciStarter. The method first generates a list of all popular projects in SciStarter by ranking all projects based on the number of users who contributed to them. Then, if the project requiring explanation is included in this top N list, popularity may be used as the explanation method. In such a case, the explanation will state that this recommended project is “Hot on SciStarter” and this may be accompanied with an indication on how popular this project is (its rank in the popular projects list)

Location Based In this explanation method, we are looking for an overlap between the recommended project’s location-polygon, if it exists, and the location of the specific user, which can be extracted from their IP address, if available. In case an overlap between the two locations if identified, we may use the location based explanation method for the recommendation explanation. In such a case, the phrase “Projects near you” will be used for the explanation.

Co-funded by the Horizon 2020
Framework Programme of the European Union

3.2.3 An Algorithm for Selecting Explanation Types

Algorithm 1 describes the algorithm for choosing an explanation type for each recommended project. The input to the algorithm is an ordered list of recommended projects outputted by the Hybrid-BPR model, as well as a set of methods for generating explanation types. The algorithm selects one explanation type for each of the recommended projects, choosing from the possible explanation types described above.

Prior work has discussed the advantage of rule mining approaches (such as association rules) in generating straight-forward explanations for users [32]. Additionally, research has shown the benefits of Content-Based and “Neighborhood-Style" based explanations in generating recommendation justifications to users [33,22].

Informed by these results, we prioritize the generation of explanation types for each recommended project in a list of recommendations (R) as follows: The algorithm first applies the Topic Association Rule (TAR) method. If a matching topic is found for the recommended project, this method is selected for explanation. Otherwise, the algorithm checks if the project location is in proximity to the user’s location based on the Location Based list (LB). If this is the case, the Location Based explanation type is chosen for the project. Otherwise, the algorithm attempts to apply the item-item or features based method (IC). If the project matches the output of both methods, the method with the highest score is selected. If the project matches the output of only one of these methods, this method is selected. If an explanation still cannot be generated, the algorithm then applies the popularity method (P) and explains the project based on it popularity in case it is one of the popular projects in the system. Otherwise, a general explanation phrasing is provided. This phrasing includes the text “Try something new”, emphasizing the fact that the recommended project is new to the user.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Algorithm 1: Assigning explanations to Hybrid BPR recommendations

Result: Ordered set of N recommended projects and the explanation type for each project

```
procedure explain(R, TAR, LB, IC, P)
    for project ∈ R do
     if project ∈ TAR then
         project.Explanation = TAR.Explanation
     else
         if project ∈ LB then
             project.Explanation = LB.Explanation
         else
             if project ∈ IC then
                 project.Explanation = max_Explanation Score(project) ∈ IC
             else
                 if project ∈ P then
                     project.Explanation = P.Explanation;
                 else
                     project.Explanation = General.Explanation
                 end
             end
         end
     end
    end
```

3.2.4 Creating Project Carousels by Explanation Types

Algorithm 1 assigns an explanation type for each of the N ordered recommendations for a given user. Take for example the following three projects recommended to a user, in decreasing order: iNaturalist (explained using the Feature-Based method), never-Home-Alone (explained using the Association rules method), and Globe at Night (explained using the Feature-Based method).

One of the requirements from SciStarter is to organize explanations in groups (or carousels [18]) in a manner that is similar to other recommendation services (e.g. Netflix). In our example, this means that the projects iNaturalist and Globe at Night would be visualized together because they belong to the same explanation type.

Importantly, this requirement may also require to re-order the original rankings obtained from the Hybrid-BPR algorithm, and visualize recommendations according to the new order. One possible reordering that satisfied the grouping requirement is to prefer larger groups over smaller groups in the visualization. Thus now ranking Globe at Night higher than iNaturalist. Another possible reordering is to prefer explanation types that work well with people such as Association Rules [33], now ranking the Globe at Night project highest among the three projects. Note that this reordering of project recommendations may affect how users perceive

Co-funded by the Horizon 2020 Framework Programme of the European Union

the recommendations since the ordering is now different from the optimization performed by the ranking algorithm. We explicitly study this in the experiment described in Section 4.3.

We tested several approaches for determining the visualization order of the project groups. The input to each approach is the ranked set of N projects grouped by explanation types. The output is an ordering over these groups that will determine how they are presented to users.

- Re-order by Leader. In this approach the order of the groups is determined by the rank of the first project in each group. Here the new order of the projects is kept as close as possible to the original ranking of the N projects.

- Re-order by max size. In this approach the order of the groups is determined by the group size, such that larger groups of projects are presented first.

- Re-order by Explanation Priority. in this approach the order of the groups is strictly determined by their explanation type, in the following order: Topics Association rules, Feature based, Location based, CF-Item-Item, popularity based, general message. The justification for this ordering directly follows from prior work. Previous studies show that most users prefer simple and short explanations [22] which are naturally generated by association rules. Second, recommendations generated with collaborative filtering approaches are less intuitive to explain compared with those generated by content-based algorithms [33]. Lastly, Shmaryahu et al. [22] showed that popularity based explanations, and general explanations result in lower user satisfaction than collaborative filtering and content-based approaches examined in a user study.

3.3 Empirical Evaluation

In this section we compare the performance of the approaches for generating recommendations described in Section 3. We will evaluate Algorithm 1 for assigning explanations in a real world study presented in Section 4.3.

3.3.1 Offline Evaluation of the Hybrid-BPR Approach

We begin by comparing the performance of the Hybrid-BPR approach to alternative state-of-the-art recommendation algorithms on SciStarter data. The evaluation was performed on historical data collected between January 2012 to May 2021. The dataset includes data from 11,223 users who interacted with 216 affiliate project (1.42 project per user on average). These projects use a dedicated API to report back to SciStarter each time a logged-in SciStarter user has interacted with the project's website or app. For our experiments, we consider users who interact with at least 2 projects, and hence remain with 4,118 users.

The alternative approaches included two collaborative filtering algorithms (CF-item-item [31] and BPR [25]) and a content-based approach, which uses the project properties to compute similarities between projects [34,35]. These approaches have been shown to provide good results in several deployed recommendation systems that used different data sets. We also used a non-personalized baseline approach that recommends projects by their overall

Co-funded by the Horizon 2020 Framework Programme of the European Union

popularity. We evaluated the prediction quality of the top-n recommendation algorithms for n = 3, 5, 7, 10 and 20 projects using the precision metric. This reflects the range of recommendations presented to the user on the SciStarter site.

We chronologically split the data using cross-set validation into train and test sets such that 10% of the latest interactions from each user are selected for the test set and the remaining 90% of the interactions are used for the train set.

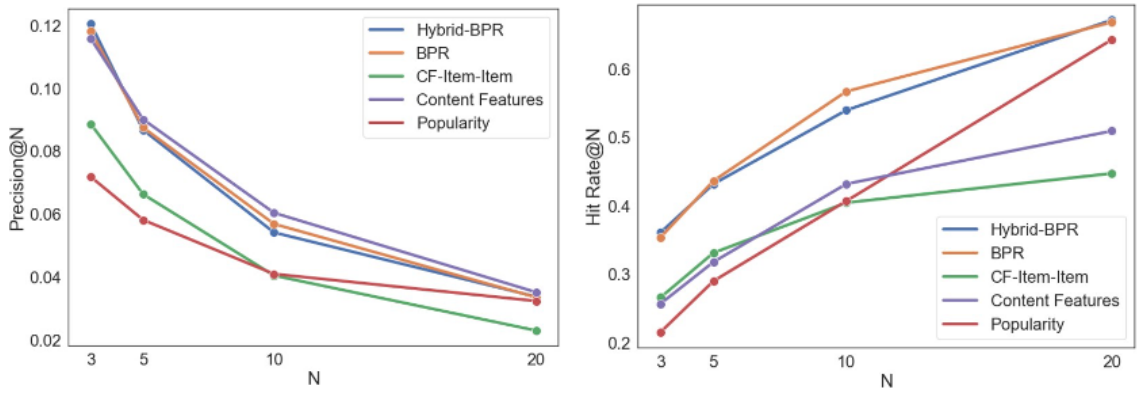

Figure 11: Precision@N and HitRate@N results of Hybrid-BPR approach on offline data

Figure 11 (left) shows the precision@N results for the five examined algorithms and for different N values of recommended projects. As can be seen from the figure, The Hybrid-BPR approach, BPR, and the content-based method obtain the highest precision. These results are significantly higher (using the Mann-Whitney test) than the popularity and the collaborative filtering item-item based approaches (there was no statistically significant difference between the performance of the top-scoring methods). The popularity and CF-Item-Item recommendation algorithms generated the lowest performance. This was especially apparent for lower values of N , which are more reasonable in the citizen science domain, as it is unlikely that users will consider to contribute to more than a handful of recommended projects. We note the overall low value of the precision metric here for all methods, which is often the case for real world applications where a user consumes only a handful of products [36,37,38,39,40].

Figure 11 (right) shows the hit rate of the different algorithms, defined as the portion of recommended lists in which users accessed at least one project that was recommended to them [41]. Naturally, the hit rate for all algorithms rises consistently as the number of recommended projects increase. The difference between the Hybrid-BPR Model and the BPR compared to all the other algorithms were statistically significant for every value of N (using the Mann-Whitney test). For N = 20 the non-personalized popularity method achieved the same hitrate score as BPR and the Hybrid-BPR method. This is because most users contributed to at least one of SciStarter's 20 popular projects.

Based on the offline results demonstrated by the Hybrid-BPR approach, we continue with this algorithm in the next steps of our methodology.

Co-funded by the Horizon 2020
Framework Programme of the European Union

3.3.2 Offline Evaluation of the Reordering Approaches

As we explained above, the SciStarter UI presents recommended items grouped together by a common feature, such as the most appropriate explanation. The order of groups results in a reordering of the recommended items, where items from the first group appear first, internally ordered by the score provided by the Hybrid-BPR recommendation algorithm, then items from the second group, and so forth. This ordering deviates from the original ordering of items dictated by the recommendation algorithm. In this section we evaluate the impact of various ordering of the projects according to the Precision@N and HitRate@N measures.
We expected that the orderings dictated by the Hybrid-BPR algorithm will perform best, and the restrictions imposed by the grouping of items will have a negative effect. We aim to measure the negative impact of every reordering approach on these metrics. In this experiment, we compared each of the re-ordering approaches to the original Hybrid-BPR list. For each user we output a list of recommendations from the Hybrid-BPR and re-order it by each of the Explanation Priority, Max Size and the Leader approaches.

Figure 12 shows the results of the precision and hit rate for n = 3, 5, 7, 10, and 20 recommended items. The metrics were computed on a train/test split of 90%/10% for every user's interactions as described in section 4.1. First, as expected, we can see that the baseline Hybrid-BPR ranking performs the best. The reordering by the Explanation Priority and Leader methods outperforms the reordering by the Size method on both metrics for every N measured, except at 20. For k = 3, the number of recommended items shown in the SciStarter UI (and also in the online study), the statistical test for HitRate@3 comparing the Explanation Priority method and the Size method is U = 2566537 with p < 0.05 (Mann-Whitney test).

Additionally, while the difference in performance between the Leader approach and the original ordering is significant, there is no statistically significant difference (Mann-Whitney test) between the Explanation Priority reordering approach and the baseline Hybrid-BPR approach. Thus, this explanation approach can be chosen without a significant decrease in the prediction performance.

Another advantage of the Explanation Priority approach is complete independence from the Hybrid-BPR approach, such that the display order of the explanation groups will remain the same even if the list of recommendations changes. In contrast, in the Leader reordering approach the order in which the explanation groups are presented to users may vary as the recommended projects change. Studies have shown [42] that having a fixed order of presentation can prevent confusion and save time for the user. Thus, Explanation Priority is our chosen reordering method to be used for our online experiment.

Finally, we study whether the different reordering approaches modify the original ranked lists that are outputted by the Hybrid-BPR approach. To measure the difference between the lists, we use the Levenshtein distance metric [43] which measures the number of edit changes between two sequences (in our case, substitutions between two ordered lists). We compute the distance between the baseline Hybrid-BPR model and the 3 reordering approaches.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 1 presents a diagonal matrix that measures Levenshtein distances between each pair of approaches, averaged over all users. As seen from the table, the Levenshtein distance from all reordering approaches to the Hybrid- BPR model is at least 13, meaning that on average, at least 6 pairs of recommendations (out of 20 recommendations) were ordered differently than in the Hybrid-BPR list. Additionally the Levenshtein distance between the different reordering approaches demonstrates that they produce different projects' re-orderings, despite their relatively similar Precision@N and HitRate@N .

Table 7: Diagonal matrix: Levenshtein distance of ordering Methods.

Method	Hybrid-BPR	Explanation Priority	Leader	Size
Hybrid-BPR	0	13.32	13.13	14.25
Explanation Priority		0	5.26	9.33
Leader			0	6.85
Size				0

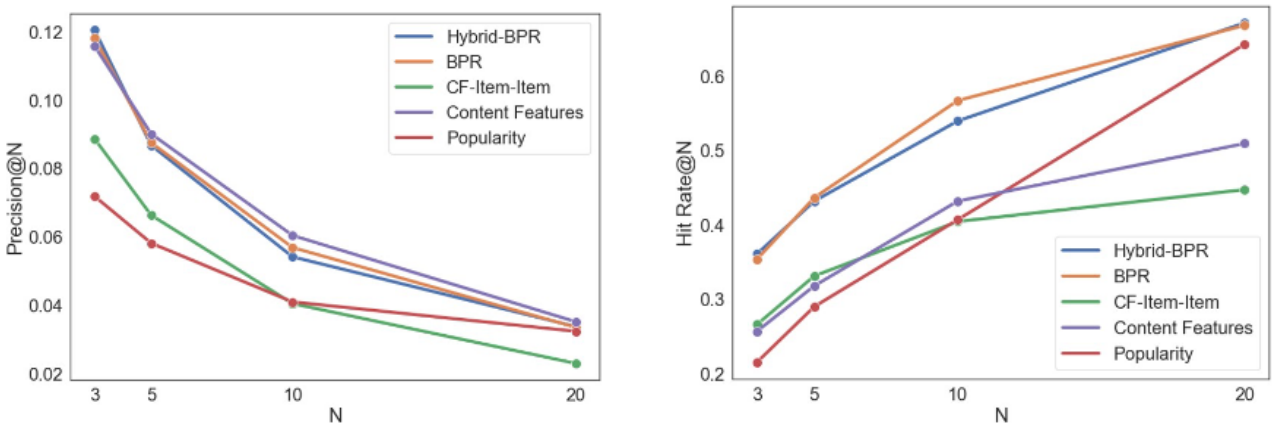

Figure 12: Precision@N and HitRate@N results of Precision approaches compared to Hybrid-BPR

3.3.3 Plan of Online Studies

In this section we report on our planned evaluation of our methodology in the real world, by conducting an online experiment on the SciStarer portal. Users who will log in to SciStarter during the experiment period will be randomly assigned to one of the two cohorts: one cohort receiving project recommendations without explanations (the Recs Cohort) and the other cohort receiving project recommendations together with explanations for each of the recommendation groups (the RecsExp Cohort). The initial ranked list of 20 recommendations

Co-funded by the Horizon 2020
Framework Programme of the European Union

for both cohorts will be generated using the Hybrid-BPR approach. For the RecsExp cohort, the recommendations will be grouped by the Explanation Priority approach.

In the experiment we will consider only users who have engaged with at least 2 projects prior to the experiment launch, to enable available data for the Hybrid-BPR model, and who have logged into SciStarter at least once in the 6 months before the experiment. Recommended projects will be displayed on the SciStarter home page in groups of 3 recommended projects. By clicking on the “See more recommendations” button , users will be able to see more recommended projects in decreasing order of relevance which will be displayed on their dashboard page. The Recs cohort will get one rolled carousel, ordered by the BPR-Hybrid and the RecsExp cohort will receive multiple carousels as shown in Figure 13.

Figure 14 shows an example of the top three recommended projects for two users, one user from the Recs cohort (top) and one user from the RecsExp cohort (bottom). As can be seen in the bottom figure, explanations are presented as the header of the recommended projects list (e.g., “Try projects with new topics” in this case, which represent the association rules explanation type). Additionally, a question mark appeared near every project name in this cohort. By hovering over this question mark, users will receive additional explanation details per recommended project (e.g., “Because you contributed to the project with Molecular and Cells topics in the past” for the Flu Near You recommendation.)

The study will be conducted in a manner that preserves privacy and transparency. A blog post announcing the new recommendation tool will be published to all SciStarter users prior to the experiment. This blog will contain detailed explanations about recommendation algorithms and about the explanations types used in the study. Additionally, users will be given the option to opt-out from receiving recommendations at any point in the experiment.

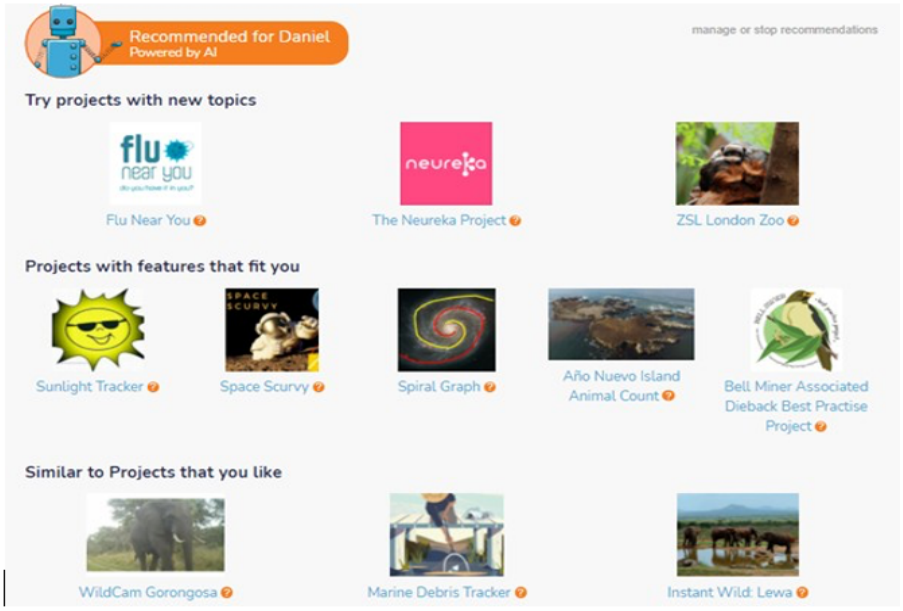

Figure 13: The SciStarer dashboard, an example of the user interface for the RecsExp cohort, showing three groups of recommendations, each with an associated explanation. Figure also shows the opt out option available to users.

Co-funded by the Horizon 2020
Framework Programme of the European Union

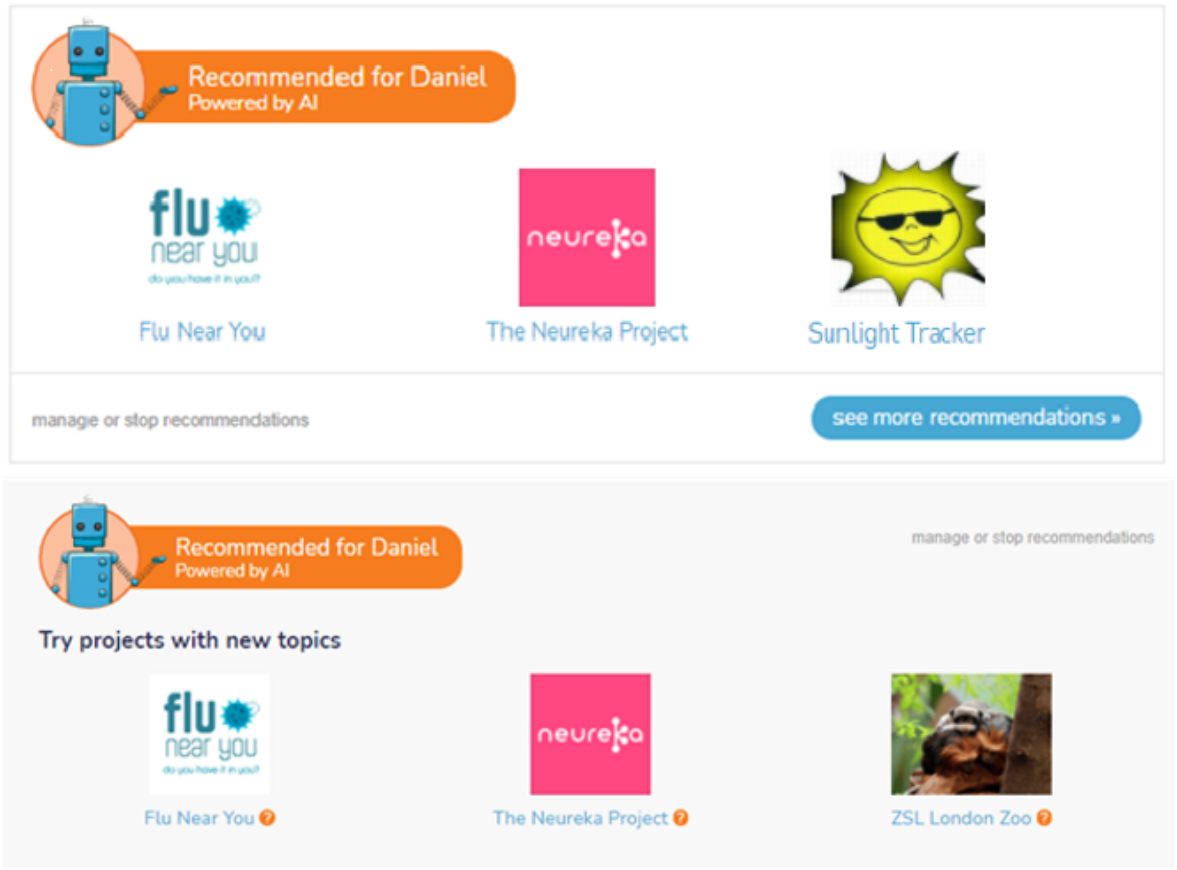

Figure 14: The SciStarer homepage, an example of user interface for the Recs cohort (top) and RecsExp (Bottom).

The results of the online experimentation, together with the results of a user study performed post the online experimentation, will be reported in a future report.

3.4 Mapping Explanation Types to WeNet's Diversity Dimensions

The WeNet project addresses diversity through the lens of social practices. Social practices are routine behaviour like going to work, cooking and showering, which integrates different kinds of elements, such as bodily activities, material artefacts, skills, and associated meaning [44]. This view of diversity ventures away from a shallow notion of demographic differences between people (based on gender, race, geographic location, and the likes) to a deeper view, which considers multiple aspects of human existence and behaviour. Specifically, the literature defines the different aspects of social practices as follow:

- ***Material*** covers all physical aspects of the performance of practice encompassing objects, infrastructures, tools, hardware including the human body.
- ***Competence*** incorporates skills, know-how, (background) knowledge as well as social and relational skill which are required to perform the practice
- ***Meaning*** incorporates understanding, beliefs, values, norms, lifestyle, emotions, and social and symbolic significances

Co-funded by the Horizon 2020
Framework Programme of the European Union

We now consider the different explanation types used in our algorithm and their relations to the Material, Competence and Meaning aspects. Table 8 presents the social practices aspects associated with each explanation type. As detailed in the table, all dynamic explanation types considered by our algorithm are covering social practices and are thus matching explanations to users based on their deep (and changing) behaviour in the system. Only if no explanation is obtained by any of these algorithms, our approach falls back to a popularity based or general based explanation. Indeed, in these two explanation types, the user's uniqueness and diversity is not taken into consideration, and a “one size fits all” solution is utilised as a last resort.

Table 8: Explanation Type Analysis – Diversity Dimensions

Explanation Type	Social Practice	Justification
Topics Association Rules	Material, Competence, Meaning	Projects associated together indicate **latent relations between topics**. These relations are the result of users' behaviour in the system which in turn are influenced by deep aspects of the users: the material they possess, their skills, and the meaning they attribute to different topics.
Content Feature Based	Competence, Meaning	Similar projects are identified using **textual relations between project metadata.** This metadata, available to the user, describes the expected effort of each project (competence) and its overall intentions (meaning).
Collaborative Filtering Item-Item	Material, Competence, Meaning	Similar projects are identified by **matching consumption patterns between users**. These patterns may be a result of any latent characteristic of users, not restricted to their demographic features.
Collaborative Filtering User-User	Material, Competence, Meaning	Similar projects are identified by **matching consumption patterns between users**. These patterns may be a result of any latent characteristic of users, not restricted to their demographic

Co-funded by the Horizon 2020 Framework Programme of the European Union

		features.
Popularity Based	Shallow	Explanation is ignoring the user's unique behaviour in the system, using instead a fall back solution of "one size fits all" solution.
General Message	None	Explanation is ignoring the user's unique behaviour in the system, using instead a general message.
Location Based	Shallow	Recommendation and explanations based on location are using a shallow demographic feature, failing to refer to the wider user's social practices.

Co-funded by the Horizon 2020
Framework Programme of the European Union

4. Conclusion

In this reporting period WP4 has engaged in a wide range of research and development activities as part of the WeNet project. These activities were conducted within the WeNet eco-system (developing technologies, conducting research, and participating in pilot1) as well as external to the WeNet system (building upon our capacity to use the external SciStarter platform and dataset).

First and foremost, we have extensively developed the Incentive Server for the project, integrating it with main components of other work packages and incorporated it into Pilot1, WeNet's leading implementation for the reporting period. We have used the pilot to run a randomised experiment comparing intectives across types and locations and have reported the results of this endeavour in this delivery. Then, we used the data collected during this pilot to develop and evaluate (in offline setting) a personalized approach for incentive design, which builds upon shallow as well as deep diversity characteristics of users.

Additionally, we have advanced our close cooperation with WP9, the ethics group in the WeNet project. We described in this report the first fruit of this cooperation, our mutual work on Transparency in Machine Generated Personalization, which developed best practices and a checklist for system designers and users of such systems. This joint work has resulted in a joint publication in the UMAP conference. We have also described our planned future work in this important joint research area.

Finally, we have described our additional work with SciStarter on developing recommendation based incentives for increasing motivation and improving productivity and engagement for single users. Specifically, we have designed adaptive explanations based on users' characteristics and behaviours in the SciStarter system. In this line of work, we built on the large scale data available in the SciStarter system. Our implementations and trials point to the potential of these approaches in designing adaptive based explanations for users in these environments.

Co-funded by the Horizon 2020
Framework Programme of the European Union

REFERENCES

[1] Slivkins, Aleksandrs. "Introduction to multi-armed bandits." *arXiv preprint arXiv:1904.07272* (2019).

[2] A. Segal, K. Gal, E. Kamar, E. Horvitz, and G. Miller. Optimizing interventions via offline policy evaluation: Studies in citizen science. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[3] Q. Wu, H. Wang, L. Hong, and Y. Shi. Returning is believing: Optimizing long-term user engagement in recommender systems. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 1927–1936. ACM, 2017.

[4] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In Proceedings of the 22nd international conference on World Wide Web. ACM, 95–106.

[5] Yanovsky, Stav, et al. "One size does not fit all: Badge behavior in q&a sites." *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 2019.

[6] Lu, Tyler, Dávid Pál, and Martin Pál. "Contextual multi-armed bandits." *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2010.

[7] Zhou, Li. "A survey on contextual multi-armed bandits." *arXiv preprint arXiv:1508.03326* (2015).

[8] Bouneffouf, Djallel, and Irina Rish. "A survey on practical applications of multi-armed and contextual bandits." *arXiv preprint arXiv:1904.10040* (2019).

[9] Mui, John, Fuhua Lin, and M. Dewan. "Multi-armed Bandit Algorithms for Adaptive Learning: A Survey." *International Conference on Artificial Intelligence in Education*. Springer, Cham, 2021

[10] Dudík, Miroslav, et al. "Doubly robust policy evaluation and optimization." *Statistical Science* 29.4 (2014): 485-511.

[11] Beygelzimer, Alina, and John Langford. "The offset tree for learning with partial labels." *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009.

[12] Ferwerda, B., Swelsen, K., & Yang, E. (2018). Explaining Content-Based Recommendations. New York, 1-24.

[13] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[14] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.

Co-funded by the Horizon 2020
Framework Programme of the European Union

[15] Bilgic, Mustafa, and Raymond J. Mooney. "Explaining recommendations: Satisfaction vs. promotion." *Beyond Personalization Workshop, IUI*. Vol. 5. 2005.

[16] Mohammed Z Al-Taie and Seifedine Kadry. 2014. Visualization of explanations in recommender systems. *Journal of Advanced Management Science Vol* 2, 2 (2014), 140–144.

[17] Saeed Amal, Mustafa Adam, Peter Brusilovsky, Einat Minkov, and Tsvi Kuflik. 2019. Enhancing explainability of social recommendation using 2D graphs and word cloud visualizations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. 21–22.

[18] Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2021. A Methodology for the Offline Evaluation of Recommender Systems in a User Interface with Multiple Carousels. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 10–15.

[19] Behnoush Abdollahi and Olfa Nasraoui. 2017. Using explainability for constrained matrix factorization. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 79–83.

[20] Peake, Georgina, and Jun Wang. "Explanation mining: Post hoc interpretability of latent factor models for recommendation systems." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.

[21] C. Musto, Marco de Gemmis, P. Lops, and G. Semeraro. 2020. Generating post hoc review-based natural language justifications for recommender systems. *User Modeling and User-Adapted Interaction* (2020), 1–45.

[22] Dorin Shmaryahu, Guy Shani, and Bracha Shapira. 2020. Post-hoc Explanations for Complex Model Recommendations using Simple Methods.. In *IntRS@ RecSys*. 26–36.

[23] Daniel Ben Zaken, Kobi Gal, Guy Shani, Avi Segal, and Darlene Cavalier. 2021. Intelligent Recommendations for Citizen Science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14693–14701.

[24] Kula, Maciej. "Metadata embeddings for user and item cold-start recommendations." *arXiv preprint arXiv:1507.08439* (2015).

[25] Yehuda Koren and Robert Bell. 2015. Advances in collaborative filtering. *Recommender systems handbook* (2015), 77–118.

[26] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[27] John S. Breese, David Heckerman, and Carl Kadie. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *UAI*. Morgan Kaufmann, 43–52.

[28] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Citeseer, 487–499.

Co-funded by the Horizon 2020
Framework Programme of the European Union

[29] Peake, Georgina, and Jun Wang. "Explanation mining: Post hoc interpretability of latent factor models for recommendation systems." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.

[30] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook* (2011), 73–105.

[31] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer, 291–324.

[32] Zhang, Yongfeng, and Xu Chen. "Explainable recommendation: A survey and new perspectives." *arXiv preprint arXiv:1804.11192* (2018).

[33] Daher, Julie, Armelle Brun, and Anne Boyer. A review on explanations in recommender systems. Diss. LORIA-Université de Lorraine, 2017.

[34] Hyung Jun Ahn. 2006. Utilizing popularity characteristics for product recommendation. *International Journal of Electronic Commerce* 11, 2 (2006), 59–80.

[35] Nirmal Jonnalagedda, Susan Gauch, Kevin Labille, and Sultan Alfarhood. 2016. Incorporating popularity in a personalized news recommender system.*PeerJ Computer Science* 2 (2016), e63.

[36] Ramazan Esmeli, Mohamed Bader-El-Den, and Hassana Abdullahi. 2020. Session similarity based approach for alleviating cold-start session problem in e-commerce for Top-N recommendations. In *12th International Joint Conference on Knowledge Discovery*. SciTePress.

[37] Asela Gunawardana, Guy Shani, and Sivan Yogev. 2021. Evaluating Recommender Systems. In *Recommender Systems Handbook (3rd edition)*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer.

[38] Qiwei Han, Inigo Martinez de Rituerto de Troya, Mengxin Ji, Manas Gaur, and Leid Zejnilovic. 2018. A collaborative filtering recommender system in primary care: Towards a trusting patient-doctor relationship. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 377–379.

[39] Jihoi Park and Kihwan Nam. 2019. Group recommender system for store product placement. *Data Mining and Knowledge Discovery* 33, 1 (2019), 204–229.

[40] Xi Shao, Guijin Tang, and Bing-Kun Bao. 2019. Personalized travel recommendation based on sentiment-aware multimodal topic model. *IEEE Access* 7 (2019), 113043–113052.

[41] Xin Wang, Yunhui Guo, and Congfu Xu. 2015. Recommendation algorithms for optimizing hit rate, user satisfaction and website revenue. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[42] Debbie Stone, Caroline Jarrett, Mark Woodroffe, and Shailey Minocha. 2005. *User interface design and evaluation*. Elsevier.

Co-funded by the Horizon 2020
Framework Programme of the European Union

[43] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.

[44] Reckwitz A. Toward a Theory of Social Practices: A Development in Culturalist Theorizing". European Journal of Social Theory. 2002;5(2):243-263. doi:10.1177/13684310222225432

Co-funded by the Horizon 2020
Framework Programme of the European Union