OXFORD

Interacting with Computers, 2024, 1–15
https://doi.org/10.1093/iwcomp/iwae008
Articles

Paintings, Not Noise—The Role of Presentation Sequence in Labeling

Merlin Knaeble[1,*] and Mario Nadj[2] and Alexander Maedche[1]

Human-Centered Systems Lab (h-lab), Karlsruhe Institute of Technology (KIT), 76133 Karlsruhe, Germany
Business & Information Systems Engineering, TU Dortmund University, 44221 Dortmund, Germany
Corresponding author: merlin.knaeble@gmail.com

Labeling is critical in creating training datasets for supervised machine learning, and is a common form of crowd work heteromation. It typically requires manual labor, is badly compensated and not infrequently bores the workers involved. Although task variety is known to drive human autonomy and intrinsic motivation, there is little research in this regard in the labeling context. Against this backdrop, we manipulate the presentation sequence of a labeling task in an online experiment and use the theoretical lens of self-determination theory to explain psychological work outcomes and work performance. We rely on 176 crowd workers contributing with group comparisons between three presentation sequences (by label, by image, random) and a mediation path analysis along the phenomena studied. Surprising among our key findings is that the task variety when sorting by label is perceived higher than when sorting by image and the random group. Naturally, one would assume that the random group would be perceived as most varied. We choose a visual metaphor to explain this phenomenon, whereas paintings offer a structured presentation of coloured pixels, as opposed to random noise.

RESEARCH HIGHLIGHTS

- A study was conducted to assess the impact of presentation sequence on labeling tasks.
- The key finding is that sorting by label is perceived as more varied than sorting by image or at random.
- This is counter-intuitive, as mathematically the random presentation sequence is most varied.
- An important implication is that labeling tasks should be designed with a non-random structure in mind.

Keywords: *interactive labeling; annotation; interactive machine learning; training data; crowd work; crowdsourcing; task design; variety; self-determination theory.*

1 Introduction

The ongoing digital transformation is commonly associated with an increase in automation, hence tasks that were previously performed by humans, are nowadays taken over by machines. However, it has also given rise to a concept called heteromation (Ekbia & Nardi, 2014), which refers to small tasks that typically can not be efficiently automatized fully. They are known to encompass motivational challenges. Self-checkout lanes in grocery stores are a form of entirely unpaid heteromation, whereas crowd work on platforms like Amazon Mechanical Turk[1] (MTurk) is often compensated insufficiently. Workers can register on platforms like MTurk to be able to take over small virtual tasks assigned by anyone willing to pay. On such platforms, a prominent task type is labeling, defined as adding additional information to existing data (Bernard *et al.*, 2018a). It is often performed with the goal to train machine learning (ML) models, which require large quantities of labeled training data (Bernard *et al.*, 2018a). However, labeling refers to a labor-intensive and error-prone process (Bernard *et al.*, 2018a, Nadj *et al.*, 2020, Zhang *et al.*, 2008) as tasks can be cut into small segments and be performed from anywhere at any time. Investigative journalists, however, have long pointed out that not enough attention is paid to the circumstances in which labeling work can become more fulfilling (Lee, 2018, Yuan, 2018). This problem is exacerbated by the fact that low-skilled workers will be even more forced to take on similar tasks, as further low-skilled jobs will be threatened by automation in the near future (Reese, 2016). Insights from interviews of labeling workers have described labeling as 'boring, repetitive, never-ending work' (Lee, 2018, para. 34) and something one would be doing 'for the money' (Yuan, 2018, para. 25) only. Self-determination theory (SDT; Deci & Ryan, 1985)—a well-established theory for understanding motivation and the role of humans in technology (Peters *et al.*, 2018, Szalma, 2014)—points at an explanation for this, whereas a lack of autonomy leads to a depletion of intrinsic motivation. It has demonstrated its strengths in different contexts, recently also for crowd work (Deci *et al.*, 2017, Durward *et al.*, 2020, Manganelli *et al.*, 2018, Van den Broeck *et al.*, 2010). SDT hereby introduces basic psychological needs such as autonomy as well as motivational qualities such as intrinsic motivation as mediators to understand the influence on psychological and work performance outcomes. Hence, an investigation based on SDT

[1] https://www.mturk.com/

Received: July 3, 2023. **Revised:** November 1, 2023. **Accepted:** November 12, 2023

seems adamant, if one is concerned with the regards of labeling workers.

Although task variety is known in SDT as a driver of human autonomy and intrinsic motivation (Deci & Ryan, 2008, Deci *et al.*, 2017), there is little research related to labeling (Durward *et al.*, 2020, Knaeble *et al.*, 2020). Hereby, task variety is defined as the perception of the range of work activities required (Durward *et al.*, 2020, Morgeson & Humphrey, 2006). As a result, scholars called for both the study of different types of task designs as well as to investigate the nomological network from task variety to psychological and performance work outcomes (i.e. its underlying mediation paths) in the context of labeling (Deci *et al.*, 2017, Durward *et al.*, 2020, van der Stappen & Funk, 2021). Against this backdrop, we see potential to influence the perception of task variety by changing the so-called *presentation sequence* of sub-tasks (e.g. Jing *et al.*, 2018, Tian *et al.*, 2007). For instance, multiple labels need assignment to a large amount of instances (e.g. images) and the order of such assignments could influence the behavior of labeling workers and change their perception of the task. However, current knowledge about the effects of presentation sequences in the labeling context is limited, and there exists no systematic investigation on how they affect task variety. To guide our investigation, we rely on SDT, by placing task variety as the observable mean of our manipulation (i.e. presentation sequence). We articulate the following research questions: *How do different presentation sequences in a labeling task impact the task variety of crowd workers? What are the impacts within the SDT-based nomological network from task variety to psychological work outcomes and work performance (i.e. its underlying mediation paths) in the context of labeling?*

To answer these research questions we conduct an online experiment on MTurk with 176 participants. The primary theoretical contribution of this study resides in linking different presentation sequences to workers' perceptions of task variety, autonomy, intrinsic motivation, as well as psychological work outcomes and work performance. A second contribution lies in the mediation path analysis on the theoretical foundation of SDT in one of the first large-scale online experiments on this topic, which offers insights from a holistic viewpoint by investigating the nomological network from task variety to psychological work outcomes and work performance in the context of labeling. Hereby, we follow previous calls for research along these avenues (Durward *et al.*, 2020, van der Stappen & Funk, 2021). Thirdly, we contribute by delivering empirical evidence to a common claim in interactive ML literature, which deals with training ML models iteratively, whereas users dislike being treated as an oracle (Amershi *et al.*, 2014). We contribute practically with the translation of our findings into implications for designing interactive labeling systems.

The remainder of this paper is structured as follows: We first provide a foundational overview of the necessary theoretical background on SDT, as well as related work. Subsequently, we present our hypotheses and our research model. Then, we present our experimental design in detail. Next, we outline our results and discuss our findings. We end with a short conclusion.

2 Self-Determination Theory

SDT is a contemporary and prominent theory of motivation (Deci & Ryan, 2000, Ryan & Deci, 2000). SDT (Ryan & Deci, 2000) has been tested and demonstrated its potential in different contexts, e.g. the educational sector (Ryan & Deci, 2000), crowdsourcing (Durward *et al.*, 2020) and the workplace (Deci *et al.*, 2017). Specifically in the context of the workplace, it has been shown to be a highly accurate model of relating user motivation to work performance (Deci *et al.*, 2017).

SDT introduces innate fundamental *psychological needs* as the building blocks of motivation and a prerequisite for work performance and positive psychological work outcomes. Besides competence and relatedness, *autonomy* is a fundamental need articulated in SDT. Autonomy has also been recognized as a central element in crowd work (Durward *et al.*, 2020). Autonomy is hereby defined as a sense of volition and psychological freedom (Ryan & Deci, 2000). For instance, whether a third party enforces compliance with a set of rules. A labeling system could allow for flexibility, thus engaging above mentioned volition, or present the user with little room for choice, limiting the users and thereby their perceived psychological freedom.

Task characteristics, such as *task variety* have been introduced in the SDT as potential workplace contexts, that directly influence above mentioned needs. Hereby, they can be need supporting or need thwarting (Deci *et al.*, 2017). Ideally, task characteristics are (regardless of being by design or inadvertently) supportive of fundamental psychological needs, by for instance giving employees freedom in how to approach their job. On the other hand, they could also thwart such needs, by rigidly controlling workers. Task variety is being positioned as the core task characteristic in crowd work (Durward *et al.*, 2020). It refers to the range of tasks required to be performed by the workers.

Motivation, has been introduced as a mediator in the SDT model (Deci *et al.*, 2017, Ryan & Deci, 2000, van Hooff & van Hooft, 2017). Hereby, the spectrum of motivational qualities is arranged by their self-determination in descending order. The authors of SDT later stated that 'the type or quality of a person's motivation would be more important than the total amount of motivation for predicting many important outcomes such as psychological health and well-being' (Deci & Ryan, 2008, p. 182). Research introduces *intrinsic motivation* as the most self-determined one, doing something because the activity is perceived as inherently interesting or enjoyable (Guay *et al.*, 2000).

In the work domain one distinguishes two types of outcomes (Deci *et al.*, 2017): First, there are *work performance* outcomes. Efficiency and effectiveness have long been framed as central to assess performance (Mouzas, 2006). Research (Tuttle & Romanowski, 1985) generally equates efficiency and effectiveness to the absolute quantity and quality performance measures originally proposed (Deci *et al.*, 2017). Labeling, as a means to an end, should always be observed through this economically-driven performance lens. Thereby, both the efficiency (number of labels per time unit) of the process and its effectiveness (proportion of correct labels) are crucial. For business purposes they have been posed as diametrically opposed constructs, which have to be balanced out for sustainable success, often in form of relative efficiency (correct labels per time unit) (Mouzas, 2006). SDT further introduces, and extant crowd work research highlights (Durward *et al.*, 2020), *psychological work outcomes*, fostering 'the healthy development and effective functioning of individuals' (Ryan & Deci, 2000, p. 74). For example, well-being is therein posed as connected to positive experiences, job commitment, development, and long-term change. Research has further investigated boredom in such work environments (van Hooff & van Hooft, 2017) and computer interaction (Baker *et al.*, 2010). Hereby, boredom is defined as a 'profound negative [...] and deactivating [...] emotion' (van Hooff & van Hooft, 2017, p. 133]), which could be evoked by activities. Imagine being a labeler that is tasked with assigning a certain set of labels to hundreds of images. Boredom could easily arise form this lengthy,

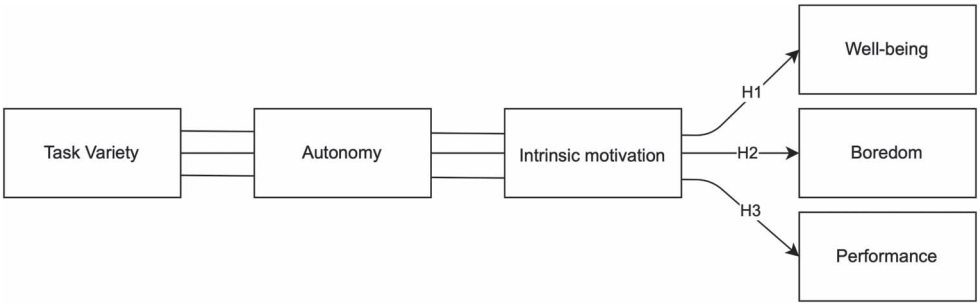

Fig. 1. Research model with hypotheses on the basis of SDT.

repetitive task. Prolonged cases of labeling have been framed as affecting all of the above outcomes (Bernard *et al.*, 2018a, Cakmak *et al.*, 2010, Zhang *et al.*, 2008), therefore their investigation seems relevant. Our research model, which we will describe in more detail in section 4, is subsequently based on SDT and is visualized in Figure 1.

3 Related Work

Within the domain of human-computer interaction (HCI), SDT has proven to be a motor theme for motivational design and well-being centered research (Ballou *et al.*, 2022, Hassenzahl, 2008). On this basis HCI research has frequently been applied SDT to the domain of games (Tyack & Mekler, 2020), gamification and learning (Lamprinou & Paraskeva, 2015). Supporting, or not thwarting, the fundamental need for autonomy has been shown to be a key determinant in video game design (Deterding, 2016), with need satisfaction issues being amplified by real-life crises like the COVID-19 pandemic reinforcing such patterns (Ballou *et al.*, 2022). Behaviour change applications, like language learning applications, or fitness tracking have been other, well-researched voluntary technology use context for SDT (Villalobos-Zúñiga & Cherubini, 2020). SDT has further been leveraged to understand the interplay of extrinsic and intrinsic motivation in crowdsourcing (Liang *et al.*, 2018), and also for human-artificial intelligence interaction design (De Vreede *et al.*, 2021).

Our research lies at the intersection of the fields of crowd work and labeling. As the predominantly used, supervised ML approaches need data to learn from, they require manual 'assignment of labels y to given input instances x' (Bernard *et al.*, 2018b, p. 1189), called labeling. Scholars in the field of interactive labeling, striving to address annoyance (Amershi *et al.*, 2014, Cakmak *et al.*, 2010) and frustration (Bernard *et al.*, 2018a, Zhang *et al.*, 2008) within the assigned workforce called for large-scale experiments (van der Stappen & Funk, 2021) and to investigate consequences on potential trade-offs for designing labeling systems (Knaeble *et al.*, 2020). However, labeling design has largely focused on performance outcomes, with a recent literature review of the field showing that while 85% of published articles on interactive labeling systems report efficiency or effectiveness as their dependent variable of choice, only 35% include user-related measures, such as well-being, motivation or boredom (Knaeble *et al.*, 2023). However, recent HCI research has shown, that the design of labeling systems is still not understood well (Zhang *et al.*, 2022). Current design knowledge in interactive labeling systems on the effects of different presentation sequences is poor, and while Jing *et al.*, (2018) or Tian *et al.* (2007) point out potential performance benefits of changing the presentation sequence, they do not study other user outcomes. Similarly, crowd work task allocation optimization research has regularly focused on such performance outcomes (e.g. Liu *et al.*, 2016). There is hence a research gap in systematic user studies on psychological work outcomes of labeling tasks in this regard.

A recent call to investigate psychological work outcomes from the research field of crowd work (Durward *et al.*, 2020) integrates well with the research gap from the field of labeling systems, as many of the tasks on crowd platforms such as MTurk are labeling tasks. Research (Ma *et al.*, 2018) already recognizes the challenging environment crowd workers are facing regarding working conditions (Asdecker & Zirkelbach, 2020, Ma *et al.*, 2018), payment (Durward *et al.*, 2020, Qiao *et al.*, 2021) and treatment by their employers (Liang *et al.*, 2015). Ekbia & Nardi (2014) point out that while crowdworkers do receive, albeit a small, monetary compensation for the microtasks they perform at the benefit of their employers, there is little, as they call it, affective reward. Such affective rewards, however, have already been connected to motivation (Deci *et al.*, 2017, Gagné *et al.*, 2022b).

Moreover, previous investigations (Leimeister *et al.*, 2009) have shown what a central role motivation plays in affecting behaviour in crowdsourcing tasks. Hereby, crowd work research on monetary incentives focuses on the study of performance outcomes (e.g. Qiao *et al.*, 2021), while typically disregarding psychological work outcomes. Although monetary incentives form the basis for achieving psychological benefits of crowd work (Durward *et al.*, 2020), they seem not enough to sufficiently motivate crowd workers. Much rather, adequate levels of compensation seem to form a fundament for other task characteristics to become effective (Durward *et al.*, 2020). Specifically, SDT-based research (Deci *et al.*, 2017) broadly connects task characteristics such as task variety and the need for autonomy to motivational qualities. In this light, the necessity of employer provided incentives for motivation (i.e. via the system design) was identified as important (Leimeister *et al.*, 2009). For instance, systems that support a range of tasks and/or a higher level of autonomy in completing those tasks are more likely to be explored and frequently used (Liang *et al.*, 2015). Thus, task variety and autonomy have emerged as relevant topics of inquiry, which are also of concern for crowd work and are studied in both virtual (Durward *et al.*, 2020) and physical setups (Asdecker & Zirkelbach, 2020). The assignment of tasks to crowd workers has already been studied in the context of skill-based systems. Thereby, matching tasks to crowd workers on the basis of cognitive skill testing shows significant increases in task performance (Hettiachchi *et al.*, 2020). Such work falls into the broader domain of algorithmic management. Oftentimes, similarly positive performance outcomes can be achieved, leading to a wide adoption of algorithmic management practices in various industries. However, algorithmic management also causes negative motivational effects in the workers subjugated to it (Gagné *et al.*, 2022a). Current research calls for a more integrated and holistic view that also

considers motivational quality, work performance and psychological work outcomes (Durward *et al.*, 2020). We therefore see a second research gap emerging, namely the need to study this nomological network of aforementioned variables in the labeling context.

4 Research Model

In Figure 1, we present our research model. Following SDT, we state three multi-layer mediation hypotheses whereas autonomy and intrinsic motivation mediate the influence of task variety on well-being (positive, H1), boredom (negative, H2) and work performance (positive, H3). We base our research on the model of SDT in the workplace (Deci *et al.*, 2017). Thereby, while the psychological needs form the foundation for human motivation in general, the needs themselves are influenced by the work environment, which may be need supporting or need thwarting. Specifically, SDT research has identified task variety as a generally positive influence (Deci *et al.*, 2017, Morgeson & Humphrey, 2006) on need satisfaction. In the context of crowd work, both autonomy (as the key psychological need of consideration) and task variety have been investigated (Durward *et al.*, 2020) independently from each other. However, there has been a lack of regard for their potential positive interrelation, as theorized in SDT. We thereby follow recent calls for research to focus our hypotheses on such multi-layer mediation effects on subsequent outcomes (Durward *et al.*, 2020)

Likewise, we assume such a positive effect to be also present in the labeling context. In particular, labeling workers typically need to assign multiple labels to a large number of instances (e.g. images), and the degree of variety in this assignment process manipulated by, for instance, the presentation sequence, could affect their autonomy in the underlying task. Consequently, the fulfillment of the psychological need of autonomy is posed in SDT as a key antecedent for human motivation (Deci *et al.*, 2017). This sense of volition and psychological freedom is, according to SDT, now moving the 'perceived locus of causality' (Deci & Ryan, 1985) and has been extensively researched in educational settings (Deci *et al.*, 1981, Flink *et al.*, 1990, Ryan & Grolnick, 1986) showing how intrinsic motivation is positively affected. More recently, similar findings have been made in the work context (van Hooff & van Hooft, 2017). For the case of labeling, imagine a worker being less tightly controlled, with not so much fine-granular task decomposition. Moreover, well-being has been positively related to autonomy in a variety of domains, from the social life, via education, to work (Milyavskaya & Koestner, 2011). Hereby, our research generally refers to mental well-being, defined as subjective positive mental health and psychological functioning (Stewart-Brown *et al.*, 2009). On this basis, and following research from the work context (Deci *et al.*, 2017), we propose that the positive effect of labeling workers' autonomy on their well-being is mediated by their intrinsic motivation. Higher levels of motivational quality, SDT also refers to these as more self-determined motivations, are thereby positively affected by autonomy. This goes directly in line with the factual definition of autonomy that presupposes a certain amount of psychological freedom. This freedom allows for positive motivational effects that in turn are the foundation for well-being. On this basis, we state:

H1: The positive effect of the labeling workers' task variety on their well-being is mediated by their autonomy and intrinsic motivation.

Research in the work context has found intrinsic motivation to be negatively related to boredom. Specifically, to 'experience self-endorsement or volition in one's actions' (van Hooff & van Hooft, 2017, p. 135) in the workplace is shown to be detrimental to the occurrence of work-related boredom. Hereby, such a sense of volition is exactly as what positive motivational qualities, and foremost intrinsic motivation, are defined as. Specifically, imagine being a crowd worker who has to assign labels to hundreds, perhaps thousands, of images. Such a monotonous task usually leads to boredom. This problem of being bored is often exacerbated by the fact that labeling workers experience a low level of autonomy, meaning they feel constantly controlled and limited in their capabilities by the system they are working with. Here, workers' perceptions of the task variety of such labeling tasks could have an influence on the subsequent constructs such as autonomy, intrinsic motivation, which may ultimately decrease boredom. Professional labeling workers state that they are doing it 'for the money' (Yuan, 2018, para. 25). Often, however, as also pointed out by investigative journalists, even money cannot prevent high employee turnover rates, underscoring the need for higher levels of autonomy and intrinsic motivation (Lee, 2018, Yuan, 2018). Following research on work-related boredom and integrating it into the larger framework of the SDT, we hypothesize:

H2: The negative effect of the labeling workers' task variety on their boredom is mediated by their autonomy and intrinsic motivation.

The workplace context may either support or thwart the psychological needs of the workers (Deci *et al.*, 2017). In turn, not only psychological outcomes are affected, but also work performance. SDT postulates intrinsic motivation to affect performance positively, which analogous to H1 and H2, seems to be driven by task variety. This assumption naturally transfers to the labeling context. However, imagine, again, performing a monotonous task like labeling images. This tedious task, combined with a potentially onerous system design due to a lack of autonomy provided, could severely compromise labeling workers' intrinsic motivation, for instance by concentrating on financial performance rewards. Hereby, the typical payment model of crowdworking, where there is a flat payment per task performed (often called pay-per-click), naturally incentivizes high efficiency. If a worker takes less time per task performed, their net hourly wage goes up. However, there have been a series of research showing that financial incentives, while almost always limiting creativity, sometimes lead people to perform worse on fairly routine and uninteresting tasks (Kohn, 1993). Common illustrations include the remembering and differentiation of similar patterns. Finally, SDT research not only frames efficiency as a key performance construct, but also effectiveness (Deci *et al.*, 2017). This goes in line with organizational research, whereas both play a crucial role in work performance (Mouzas, 2006). For labeling tasks, this goes doubly so. The data that is being produced therein is not a means of itself. Much rather it is being used further, most often for training ML models. Models trained on fundamentally incorrect data produce factually wrong results themselves—but report high accuracies, as the data they learn from and as such their perception of 'truth' is inherently flawed. Hereby, we need to also investigate changes to effectiveness of our labeling workers. They need to produce high quality output, in order for the resulting interactive labeling system to be applicable to practice. This is being regarded in the fundamental design of modern crowdworking platforms. In this study, we thereby draw upon relative efficiency as the quotient of the above for our performance investigation (Mouzas, 2006) instead of

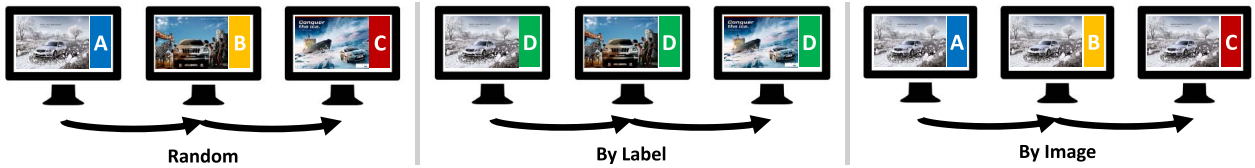

Fig. 2. Experimental treatments representing possible sequences of presentation in labeling tasks.

a disjoint view on the two performance measures. Psychological research from the SDT domain draws such an initial motivational connection to performance effects (Deci & Ryan, 2000). We surmise:

H3: The positive effect of the labeling workers' task variety on their performance is mediated by their autonomy and intrinsic motivation.

5 Research Method

We developed a fully functional labeling system providing three alternative presentation sequences (by label, by image, random) as explained in detail in subsection 5.2. For a simplified visualization of these sequences, see Figure 2. In our experiment it is served to the labelers on MTurk as a modern web application, without the need to install anything locally. Using this artifact, we conduct an online experiment. Research has shown experiments to be conducted online to be as valid as those conducted in the lab (Thomas & Clifford, 2017). They find no general issues in interaction, attention or data quality. For our specific research interest, there is another strong argument in favor of MTurk: it is where labeling is actually performed for many real-world applications in research and practice. Researchers found the vast majority of tasks on MTurk to be content creation tasks that encompass transcription, tagging and labeling (Hara *et al.*, 2018). We randomly assign the crowd workers to the different treatments. Compensation is a flat payment of USD 9.15, as to not distort any motivational variables (Gneezy & Rustichini, 2000, Ryan & Deci, 2000), following recommendations for the average compensation on MTurk (Saito *et al.*, 2019). This mirrors typical labeling task compensation on MTurk, whereas labelers are paid a flat fee for an assignment.

5.1 Domain and Task

For setting the suggested experiment in a context, we chose the marketing domain, and within that, the analysis of emotionality in car advertisements. Hereby, aforementioned cues are such 'emotion-laden' parts of advertisements 'likely to affect behavior' (Chandy *et al.*, 2001). In the context of a car advertisement, think of, for instance, a well dressed man at the wheel of the newest Audi—conveying the social status the brand wants you to associate with itself. As the labels are not mutually exclusive, any combination can occur for any given image. Such data has already been established as a noteworthy context for interactive labeling due to aforementioned high value of human input (Chen *et al.*, 2018). But apart from this initial investigation, to the best of our knowledge, there is no research on interactive labeling in this context.

The following four cues were previously identified in advertising literature and were used, along with their definitions, in our online experiment:

Firstly, **human relations** is defined as the advertisement visually displaying human interaction or relations, like families, couples and friends (Javalgi *et al.*, 1995, Xiao & Ding, 2014). We refer to **social status** as the visual display of elements of status like expensive clothes or hobbies, as well as typical signs of status such as cigars, horses and jewellery (Renson & Careel, 1986). **Joy** is known as the ad visually displaying emotions of joy. For instance, people are happy while driving the car, children excited for family holidays (Tellis *et al.*, 2019). Lastly, **freedom & mobility** is defined as visually conveying a feeling of freedom and/or mobility to the reader by, for example, showing the car driving in exotic, prestigious or hard to reach places (Sheller, 2004).

The data set for the suggested study was acquired from the magazine *The Economist* and consists of 4.617 scanned print-advertisements. From this, we randomly selected 44 images to be labeled during the experiment. The participants need to label a set of advertisement images with the previously mentioned cues. Depending on the experimental treatment, the actual process of labeling will differ. However, the final result will be the same across all participants: all four cues are labeled for the same set of images.

5.2 Treatments

We employed a three-treatment, three-group between-subject experimental design. This allows us to extend the duration of the time each participant spends labeling, without stretching the experiment duration to an unfeasible length for the participants to attend. Furthermore, it allows for using the same data set across all design configurations, respectively treatments, as they are executed by different participants. Hence, comparability of the results is guaranteed.

On a more abstract level, consider the labeling task as a set of tuples, with one element representing the label, and the other the instance (e.g. image) that is being labeled. For example, let A, B, C, D be our labels that need to be assigned, whereas $1, 2, 3, 4$ are the instances. Now $A1$ refers to the task of checking whether label A applies to instance 1, a binary decision. To obtain a complete coverage of our exemplary data set, we need to check all combinations of labels and instances. For arranging such tuples one may sort by either of the two parts, or refrain from ordering them at all. This results in our three treatments. The first approach is to sort *by label*, or in terms of our abstract example $A1, A2, A3, A4, B1, B2, \ldots D3, D4$. Note that we fix the label, until all images for that specific label have been processed. In this regard, the order of the images within each label is negligible. Another approach would be sorting *by image*, the exact opposite: $A1, B1, C1, D1, A2, \ldots, D4$. Now, the order of the labels is subordinate to that of the images. Lastly, one could disregard this orderly fashion altogether and apply *random* sorting: e.g. $B2, C4, A1, B1, D3, A4, \ldots$. Figure 2 shows these treatments in the context of our task.

5.3 Experimental Artifact

To enable the participants to fulfill their task and to investigate effects in a scenario that is comparable to real-world labelers, we developed a fully functional labeling artifact that can be used in the online experiment. The artifact has merit for future extension or real-world use. It is served to the labelers on MTurk as a modern

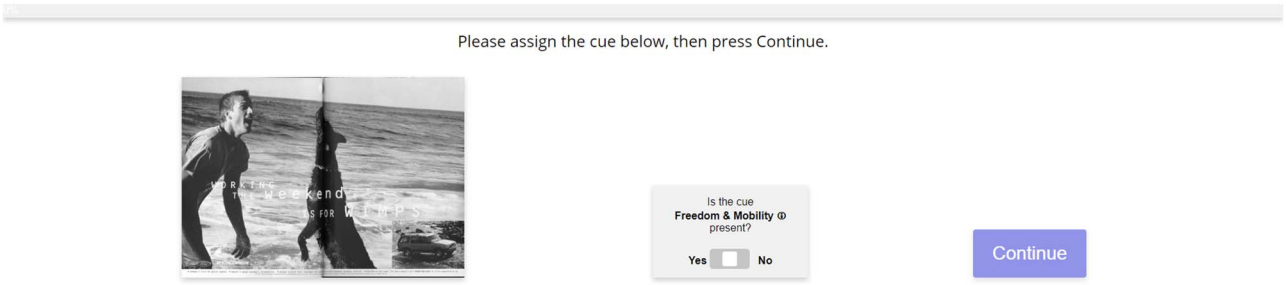

Fig. 3. Artifact (base view).

Fig. 4. Artifact (zoom functionality).

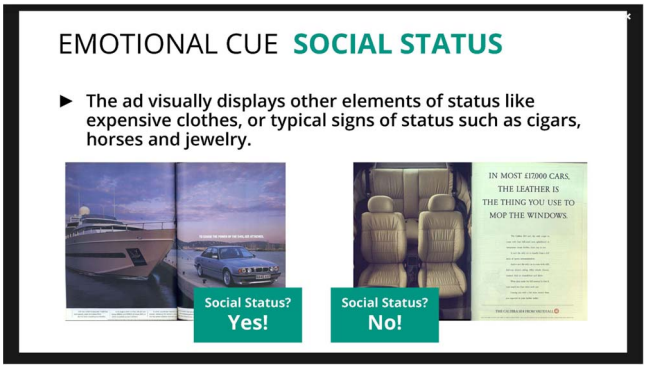

Fig. 5. Artifact (help functionality).

web application, without the need to install anything locally. It is tightly integrated with our experimental onboarding and training procedures, as well as the exit questionnaire.

As depicted in Figure 3, the artifact consists of a user interface that displays the currently selected image on the left side, and, on the right, the cue that is to be labeled. On the very right the user finds a button to continue to the next step, once they are confident they have labeled everything correctly. As not to incite the user towards selecting yes or no for any given cue, the cues are on a neutral state by default.

If the cue selector on the screen remains in a neutral position the button to continue to the next image is disabled, as can bee seen in Figure 3. Only if the switch is in a non-neutral state (Yes or No) the user can continue.

Our labeling artifact offers two further features: a zoom and a help functionality. The user may want to take a closer look at a an image. If they click on it, it is shown larger (Figure 4). The user can click anywhere to go back. While we thoroughly brief the users on the definition of the cues (human relations, social status, joy, freedom & mobility) and establish a quiz challenging their knowledge, it may happen that a user forgets the definition of a certain cue or is uncertain of its exact specification. Each of the questions is paired with a set of answer options, only one of which is correct. If the participant selects the right answer they are congratulated. For a wrong answer, they are displayed what the correct choice would have been. Furthermore, they are offered to recap, by watching the appropriate part of the introduction video again. Therefore, we implemented a help functionality. If the user clicks on a cue name, the definition of the cue along with a positive and a negative example (just as in the introduction video) is shown (Figure 5). Again, clicking anywhere takes the user back.

5.4 Participants and Procedure

Using G*Power (Faul *et al.*, 2007, 2009), we performed a statistical power analysis to estimate the required sample size for a multivariate analysis of variance (MANOVA). We followed standard assumptions and prerequisites: $f^2 = 0.0625$ (small to medium effect size; Cohen (2020)), $\alpha = 0.05$, $1 - \beta = 0.95$ and three groups.

On this basis, the power analysis computes a minimum required sample size of 171. To add enough buffer for exclusions, due to willful manipulation, technical errors or interruptions, we invited a total of 200 participants. The participants are compensated via a flat payment, as to not distort any motivational variables (Gneezy & Rustichini, 2000, Ryan & Deci, 2000). We followed recommendations for the average hourly compensation on MTurk (Saito *et al.*, 2019) and payed a fee of $9.15. Hereby, we create the prerequisite of an adequate payment, that fosters the presence of psychological benefits (Durward *et al.*, 2020). To limit cultural influences in the task of emotional cue recognition, we only invited participants from the United States. Furthermore, we only allowed *MTurk Master* qualified workers, which sets a minimum amount of tasks fulfilled and an acceptance rate of 95% or higher.

Participants were asked to report their first language, gender, age and highest education. Only one participant reported a different first language from English (Russian). 94 participants reported their gender as male, 82 as female, none as diverse. Participants are between 18 and 73 years old, with the average participant being 41.22 years old ($\sigma = 10.57$). One participant reported an education below a high school equivalent, 80 participants reported high school degrees and 76 bachelor's degrees. 14 reported master's and five doctoral degrees.

The duration of the experiment is approximately one hour, consisting of the following sections: the introduction (5 min), a training phase (10 min), the labeling interaction (25 min) and a final questionnaire (15 min).

The introduction consists of a video, giving general information about the experiment. After the video has played, the participants have to answer a short quiz to confirm whether they understood what is expected of them. In the training phase, the participants are shown a video that explains the artifact and their task. Due to

the between-participant experimental design, the participants are only shown the video corresponding to their group. Afterwards, they can try out the artifact themselves, for a few example images. Then the interaction phase starts, in which the participants' performance is recorded. After completing all data points that need to be labeled, the participants are presented with a final questionnaire.

5.5 Measures

All measurements except the objective performance measure are performed in the exit survey of the experiment. We use self-reported measures evaluated on a seven-point Likert or semantic differential scale. We provide a detailed overview of all survey items in the appendix.

5.5.1 *Subjective Measures*

To assess task variety, we used the appropriate items from the Work Design Questionnaire (WDQ) (Morgeson & Humphrey, 2006). The WDQ has already been successfully applied in diverse contexts like work and private life, to investigate the perceptions of technology designs (Fujimoto *et al.*, 2016). To assess autonomy, we applied the appropriate items from the Basic Psychological Need Satisfaction and Frustration Scale (BPNSFS) (Chen, 2015), in its version for the work domain (BPNSFS-WD) (Schultz *et al.*, 2015). Research previously recommended the usage of these measures for assessment in human-technology interaction (Szalma, 2014). For measuring the intrinsic motivation, we followed the recommendations of the Situational Motivation Scale (SIMS) (Guay *et al.*, 2000). Well-being is established as a psychological work outcome in SDT (Deci *et al.*, 2017). We measured well-being with the Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS) (Stewart-Brown *et al.*, 2009). Additionally, we included boredom, following the recent investigations on work-technology-induced boredom (Szalma, 2014) and leveraged the established 'boredom questionnaire' (van Hooff & van Hooft, 2017) to measure it.

5.5.2 *Objective Measures*

Work performance is measured objectively, without the involvement of the exit survey. Relative efficiency as our overall measure for performance is then rated as effectiveness (correct labels) divided by efficiency (time needed), following Mouzas (2006) as a balanced performance measure in the work context. To assess effectiveness, we measure the accuracy of the labeled items. Whether the answers are correct, is determined against a gold standard labeled instance of our data set. We derived this ground truth based on full inter-coder agreement between two doctoral researchers in marketing, one marketing graduate student and the first author of this article. The former three have extensive knowledge on (car) advertisements and emotional cues, while the latter represents about the level of knowledge of a participant, being trained by a short introduction to the topic, containing examples and definitions. As each participant is required to label the same images, this seems appropriate to measure participant effectiveness. Furthermore, we propose to assess efficiency as the time needed to complete the task. When measuring timings in the labeling system, special care needs to be taken considering loading times. In an online experiment, users may be faced with different internet connection speeds. To alleviate the influence of this on the task, the timing for the individual steps is only started after the site (and foremost the contained images) has loaded completely. Time was stopped when the 'Continue'-button was pressed. Relative efficiency as our performance measure is thereby the quotient of the two.

6 Results

6.1 Attention Checks

Much alike laboratory experiments, online experiments can be jeopardized by inattentive participants or outright exploitative behaviour (Thomas & Clifford, 2017). Furthermore, platforms like MTurk run the novel risk of fully automated bots answering the survey in such an intelligent fashion, that it is almost unrecognizable for the researcher. To mitigate both issues, we propose to employ a set of attention checks. Also called screeners, they are 'methods for catching and removing problematic respondents' (Thomas & Clifford, 2017, p. 186). Traditionally this is done with items in a survey, closely resembling trick questions. We added established checks in the questionnaire such as 'Please check the third box from the right'. Also, we have the results from the comprehension quiz to account for. As smart bots may already be able to circumvent such challenges, we imposed additional hurdles. The button to continue to the next step is hidden from view until the introduction video has finished playing. The human user sees it appearing afterwards, but for the bot, operating on the source code of the web page, the button is present at all times. As the timings for the individual steps are recorded, too short of a stay on the video step indicates manipulative behaviour of a human or the presence of a bot. Lastly, we added a mandatory free text field into the questionnaire. The participant is asked to enter the 'abbreviation of the research institution the experiment is from', with the hint that the logo in the top left of the survey shows those letters. We conducted the experiment with 200 participants. From those, 192 passed all attention checks and were permitted for further analysis.

6.2 Factor Analysis

To check the consistency and validity of our measurement model, we employed a confirmatory factor analysis (CFA) using *lavaan* (Rosseel, 2012). Following established approaches (Hair, 2019), we first excluded items with indicator loadings smaller than 0.7. We could confirm that all constructs have a higher average variance extracted than 0.5. We find the coefficients ω (following Bentler, 1972, Bentler, 2008, Bollen, 1980, McDonald, 1999) and Cronbachs α (Cronbach, 1951) to be larger than 0.7. Hence, we can confirm the convergent validity. The Fornell–Larcker criterion is met (Fornell & Larcker, 1981) and we find the heterotrait–monotrait ratio to be below 0.9. This further confirms the discriminant validity. We report a χ^2 to degrees of freedom ratio of 1.687, which is well below the threshold of 3. Our root mean square error of approximation is 0.062, hence below 0.08. The data meets the requirement ($\geq$ 0.9) for the comparative fit index at 0.903. Thus, we report meeting common standards for the goodness of fit. Our appendix offers further details in two tables. Concluding, the data fulfills all required criteria of a CFA evaluation.

6.3 Hypothesis Testing

Based on our research model, we analyzed the underlying data. As with the factor analysis, we relied on *lavaan* (Rosseel, 2012). We report significant ($p < 0.05$) and trend-level ($p < 0.1$) findings (following e.g. Findlater & McGrenere, 2010, Orzech *et al.*, 2016).

We find a series of significant direct effects in line with SDT, visualized in Figure 6. We report a significant ($p < 0.001$, $\beta = 0.288$) positive, direct effect between workers' perceptions of task variety and autonomy. Thus, the primal effect between task characteristics and psychological needs, as described in SDT, hereby holds for labeling workers, as well. The investigation of such links is highly important for labeling system design, especially in crowd work,

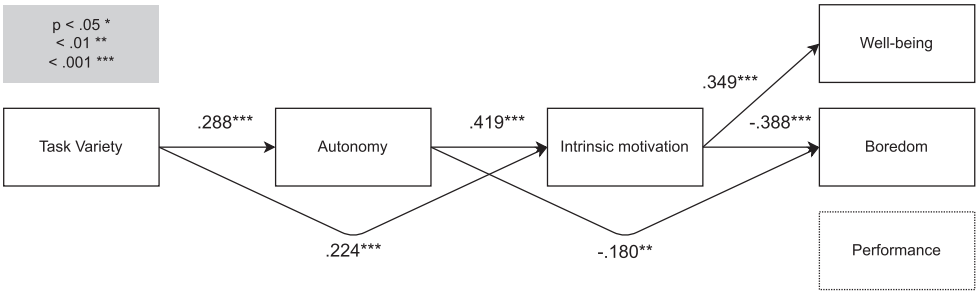

Fig. 6. Significant direct effects.

Table 1. Results of the Hypothesis Tests (p-values, β-coefficients, support indicated by ✓, no support by ✗).

		p	β	
H1	T.V. $\rightarrow \cdots \overset{+}{\rightarrow}$ W.-b.	0.009	0.042	✓
H2	T.V. $\rightarrow \cdots \overset{-}{\rightarrow}$ Bor.	0.005	-0.047	✓
H3	T.V. $\rightarrow \cdots \overset{+}{\rightarrow}$ Per.	0.618	0.108	✗

Table 2. Means (μ) and standard deviations (σ) for each treatment group.

	By image		By label		Random	
	μ	σ	μ	σ	μ	σ
Task Variety	0.388	0.255	0.459	0.239	0.359	0.254
Autonomy	0.650	0.210	0.572	0.277	0.509	0.273
Intrinsic m.	0.671	0.231	0.700	0.224	0.629	0.251
Well-being	0.697	0.209	0.714	0.202	0.739	0.235
Boredom	0.212	0.186	0.198	0.221	0.276	0.264
Performance	14.408	4.644	13.966	4.431	12.548	4.238

as the task characteristics could be adjusted in such scenarios. Along the chain of effects proposed by SDT we further find significant positive, direct effects between autonomy and intrinsic motivation ($p < 0.001$, $\beta = 0.419$), as well as intrinsic motivation and well-being ($p < 0.001$, $\beta = 0.349$). We find a significant negative direct effect of intrinsic motivation on boredom ($p < 0.001$, $\beta = -0.388$). Additionally, we find significant direct effects between task variety and intrinsic motivation (positive, $p < 0.001$, $\beta = 0.224$), as well as autonomy and boredom (negative, $p = 0.004$, $\beta = -0.180$).

In terms of the nomological network, i.e. indirect effects, we find strong support for our hypotheses H1 and H2, regarding well-being and boredom, respectively. The multi-layer mediations are significant at $p_{H1} = 0.009$ and $p_{H2} = 0.005$, with $\beta_{H1} = 0.042$ and $\beta_{H2} = -0.047$, which extends previous research on task variety (Durward *et al.*, 2020) beyond mediators into psychological work outcomes (cf. Table 1). Research on the basis of SDT has previously shown the beneficial effects of task variety, e.g. in physical exercise education, where regularly changing out training routines positively affects motivation and subsequent outcomes (Dimmock *et al.*, 2013). In labeling work, we could show similar benefits by merely changing the presentation sequence. However, we find no such effect for work performance ($p_{H3} = 0.618$, $\beta_{H3} = 0.108$). A probable explanation for this lies in the rather limited experimental time frame of one hour. Prolonged exposition to the treatment could have revealed such effects, as already illustrated by SDT research on long-term implications in other domains such as education (Ntoumanis & Standage, 2009). Regarding the control variables we report some significant direct effects, explained in detail in the appendix.

6.4 MANOVA

We analyzed the experimental data with established statistical methods. For our type of data, a MANOVA is appropriate. We meet the assumptions of multiple continuous dependent variables, and multiple independent groups. Furthermore, all observations are independent (between-subject design). Our sample size is adequate. Using standard profiles, we excluded extreme univariate ($3 \cdot IQR$ Tukey-Test) and multivariate (Mahalanobis distance ≥ 3) outliers according to Hair (2019). Hereby, we excluded another 16 participants, to obtain a final valid sample size of 176. To check for multicollinearity, we analyzed pairwise Pearson correlations. Both in the entire data set and in each treatment group the correlations were smaller than 0.7. However, many of the observed constructs return non-normally distributed data. Hence, the data fails the assumption of multivariate normality. We use a non-parametric equivalent (Dobler *et al.*, 2018, 2020) of MANOVA (*rankMANOVA*), to deal with the issue of non-normality. Subsequently, all post-hoc tests are also non-parametric. Specifically, their univariate post-hoc tests are already adjusted for Alpha error inflation, hence an (e.g. Bonferroni) correction is not required.

The non-parametric MANOVA finds a significant effect of the presentation sequence ($p = 0.022$). Hereby we followed Dobler *et al.* (2020) to test the overall effect of the presentation sequence on task variety, autonomy, intrinsic motivation, well-being, boredom, and performance. The subsequent pairwise comparison post-hoc tests are shown in Table 3, following the procedure set out by Dobler *et al.* (2020) and followed by e.g. Benke *et al.* (2022), Coll *et al.* (2020), or Haug *et al.* (2023). We find multivariate p-values for the comparison by label vs. by image (L-I) of $p_{L-I} = 0.041$, for random vs. by image (R-I) of $p_{R-I} = 0.156$ and for random vs. by label (R-L) of $p_{R-L} = 0.021$. Table 2 shows means and standard deviations for our measures. Hereby, the results for the subjective Likert and semantic differential scale measures lie on a scale of 0 (lowest) to 1 (highest). The work performance is measured objectively as correct labels per minute. We find no significant difference between the correctness of labels between the group, all are within 1 percent-point of 85%. Hence, resulting differences in performance here are due to completion time.

In terms of group comparisons, sorting by label ($\mu = 0.459$, $\sigma = 0.239$) leads to higher perceptions of task variety compared to sorting by image on the trend level ($\mu = 0.388$, $\sigma = 0.255$, $p = 0.096$, $d = 0.284$) and significantly higher than the random group ($\mu = 0.359$, $\sigma = 0.254$, $p = 0.031$, $d = 0.398$). For autonomy, we find a significant difference between random ($\mu = 0.509$, $\sigma = 0.273$) and sorting by image ($\mu = 0.650$, $\sigma = 0.210$, $p = 0.004$, $d = 0.543$), whereas the latter group reports more. We find a trend

Table 3. MANOVA pairwise p-values and Cohen's d. **Significant** effects ($p < 0.05$) are **bold**, and *trend level* ones ($p < 0.1$) in *italics*.

	L-I		R-I		R-L	
	p	d	p	d	p	d
T. V.	0.096	0.284	0.653	0.115	**0.031**	0.398
Aut.	0.133	0.298	**0.004**	0.543	0.217	0.244
I. m.	0.451	0.123	0.359	0.179	0.097	0.303
W.-b.	0.648	0.078	0.134	0.193	0.282	0.115
Bor.	0.287	0.063	0.379	0.284	0.056	0.348
Per.	0.690	0.099	**0.026**	0.415	0.072	0.316

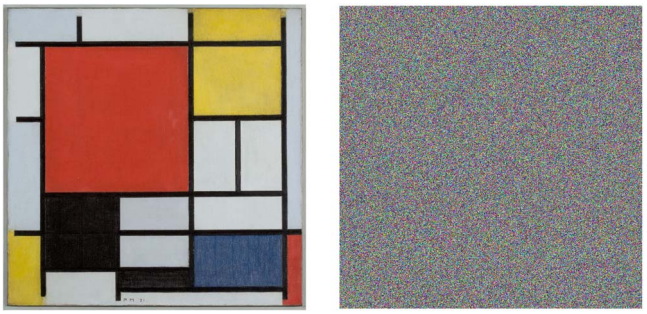

Fig. 7. Painting by Piet Mondrian (Mondriaan, 1921) on the left and Random Noise (Huiberts, 2019) on the right.

level difference for intrinsic motivation, where sorting by label ranks higher ($\mu = 0.700$, $\sigma = 0.224$) than the random approach ($\mu = 0.629$, $\sigma = 0.251$, $p = 0.097$, $d = 0.303$). We find no group differences in well-being. For boredom we find similar results to intrinsic motivation, where sorting by label retrieves a better result (i.e. lower boredom) ($\mu = 0.198$, $\sigma = 0.221$) than the random approach ($\mu = 0.276$, $\sigma = 0.264$, $p = 0.056$, $d = 0.348$) on the trend level. Regarding performance, we find a significant difference between random ($\mu = 12.548$, $\sigma = 4.238$) and sorting by image ($\mu = 14.408$, $\sigma = 4.644$, $p = 0.026$, $d = 0.415$), as well as a trend level one between random and by label ($\mu = 13.966$, $\sigma = 4.431$, $p = 0.072$, $d = 0.316$). Each time, the structured approach (by label, by image) returns higher performances than the random one.

Observing the control variables, we found no significant differences between the groups. Groups are comparably sized, with 59 samples in the by image group, 60 for the by label group and 57 for the random group.

7 Discussion

7.1 Structure Induces Variety

We find that introducing structure (by sorting by label) produces a significant difference, regarding how workers perceive task variety. This is a highly surprising result. Naturally, one would assume that the random group would be perceived as more varied than the other two groups. However, it seems that the complete disorder and intransparency of the random approach is not actually perceived as being varied. A visual metaphor for this is given by Figure 7, a famous painting by Piet Mondrian and an image consisting of randomly coloured pixels, respectively. Consider how arranging the colours into distinctly separated blocks creates a far more varied perception and an interesting composition, whereas the random (but mathematically more varied, cf. Shannon, 1948) image seems dull. Similarly, in the labeling task, changing images and labels at random seems to drown out the variety perceived by workers. In contrast, a more pronounced change after working for a longer time on the same label leads to an overall higher perception of task variety. This alone is an important distinction, structured sorting approaches can make in labeling tasks.

In historic reference, requirements for modern crowd work seem to stand somewhat in contrast to early critiques of labor force treatment. Thereby, Marx and Engels in their seminal work on means of production argue that 'for as soon as the distribution of labor comes into being, each man has a particular exclusive sphere of activity, which is forced upon him and from which he cannot escape [...]' (Marx & Engels„ 1932, part I A). Here, they argue against specialization, which their contemporaries introduced in production line factories as opposed to established trade-based manufactures. Employing a sorted sequence of presentation also represents such a sphere of exclusive activity, albeit time limited, being framed as undesirable by Marx and Engels. However, our results show the opposite is true supporting our argument of constant random changes drowning out the possibility of task variety as perceived by the workers, even though randomizing being the mathematically most varied approach.

The notion that humans perceive randomness not as a mere statistical fact and are often incapable at detecting or reproducing it, is not new (Bar-Hillel & Wagenaar, 1991, Hahn & Warren, 2009, Nickerson, 2002). However, previous research has often focused on so-called judgement ("*Were these events produced by coin flips*"?) or production ("*Create a series of coin flips.*") tasks. We extend this knowledge to the perception of randomness, as well as its effects on task variety and subsequent downstream outcomes, in the (crowd) work context. Other contexts have produced related findings, leading to randomness being drawn upon as 'an innovative design resource for supporting rich and novel user experiences' (Leong *et al.*, 2006, p. 132), with the example of randomized music playback. However, more recently, it has been uncovered that both Apples iPod, as well as the music streaming service Spotify had to move away from mathematically perfectly random arrangement of songs as users did not perceive this playback as being varied enough (Cohen, 2020). Instead, they had to introduce deliberate breaks, e.g. preventing songs from the same artist being played back to back, to prevent users from issuing complaints.

Building upon research from the domain of work design, previous research has shown how so-called *interleaved* tasks (in the form of ABCABCABC) return higher perceptions of task variety as so called *blocked* tasks (AAABBBCCC) (Derfler-Rozin *et al.*, 2016). This, however, depends on the perceived similarity of such individual tasks (A, B, C) (Pentland, 2003). Thereby, more fine-granular differences between tasks get lost in users' perceptions. Work design research has identified a potential root of such effects in the predictability of patterns in tasks (Derfler-Rozin *et al.*, 2016), however is unclear on the directionality of such effects. In our case, the two sorting parameters offered by sorting by image and sorting by label have severely different value counts. Where there are 44 images to be labeled, each image is assigned just four cues. Therefore, the sort by image pattern becomes highly predictable. After a worker has labeled just a few images, they can deduce exactly how far they have progressed for the current image. They know, how after e.g. two more cues, the image will change. In contrast, it seems far fetched to assume that workers count the number of images they have worked on in the sort by label approach. Concluding, we find a highly surprising result, in that sorting by label is perceived with a higher task variety than the random approach.

7.2 Empirical Confirmation of Previous Claims

Our findings deliver empirical support for the claims of the founders of interactive machine learning, the foundations of interactive labeling (Amershi *et al.*, 2014). Therein, and in subsequent publications (e.g. Bernard *et al.*, 2018a, Dudley & Kristensson, 2018, Knaeble *et al.*, 2023, Knaeble *et al.*, 2020, Krening & Feigh, 2019, Nadj *et al.*, 2020), it has been stated how users dislike being treated as an oracle as it is customary in active learning approaches. Active learning is a common paradigm in ML, in which the model learns from a limited set of labeled data. It then iterates over a larger set of unlabeled data, and for the instance in which it is most uncertain, the model queries an oracle, i.e. the user (Settles, 2009). From a users point of view, the black-box active learning model is equivalent to our random sorting approach. Hereby, the system requests user-input seemingly at random, without any comprehensible structure. Of course, hidden beneath the models evaluation functions, it just picks those instances in which it is uncertain. It does not pay respect to instance (e.g. image) or label based structuring towards the user. However, the user is never involved in this decision process—therefore for them, the structure could just as well have been chosen at random. This key finding allows us to support the long standing core claim of interactive machine learning, in which active learning mistreats the user, with empirical data.

7.3 Implications for SDT

Analyzing the mediation paths, we identify a tightly connected nomological net of effects in the labeling context under the theoretical lens of SDT. Hereby, the perception of task variety is fundamental to affecting autonomy as well as motivational qualities, and thereby psychological work outcomes in form of well-being and boredom. While we cannot find an effect on performance (i.e. confirm H3), this does not hinder the implications of our findings. In fact, in the single session in which the online experiment was conducted, we were able to show that we were satisfying these important needs of the workers, and thus contributing to their motivation. In this regard, work design research has already investigated the numerous positive implications of motivation and well-being on performance (Mouzas, 2006, Szalma, 2014), so we expect that the effects we observe could also be beneficial for performance, provided that a longer-term view is taken into account in the study design.

Our findings deliver strong empirical support for the basic tenets of SDT (Deci & Ryan, 2000), applied to the context of crowd work and labeling systems. Further, they are also in line with previous research on the topic. Thereby, and only after adequate payment, which we do provide, task characteristics like task variety have to be met, for downstream effects to be present (Durward *et al.*, 2020). Regarding structuring tasks, existing research has largely focused on outlining drawbacks. Most notably, Amabile (1996) has argued for task structure to limit creativity and diminish autonomy. While our task is not creative in nature, we can show that there exist tasks and contexts, in which added structure has the opposite, positive effect.

We also contribute to extending the understanding of work related technology use. From existing research, we know of the importance of task variety for motivation and performance (Hackman & Oldham, 1976, Morgeson & Humphrey, 2006). Moreover, SDT-based HCI research offers a plethora of knowledge in voluntary technology use. We do, however, find little research in the labeling domain (e.g. Durward *et al.*, 2020), which can be seen as mandatory use. With our work, we extend the application areas of SDT, and bridge knowledge previously derived from its use in the work domain and that of HCI.

7.4 Practical Implications

Against this backdrop, we could show how practitioners should focus on the sequence of the data they present their labeling workers by introducing structure into the labeling process (i.e. via by label or by image sorting). Including these findings into practical tools is of little development effort, and has strictly positive effects, as outcomes along the nomological net proposed in Figure 1 are best addressed by the structured approaches. Overall, compared to the random approach sorting by label offers advantages in terms of task variety, as well as higher intrinsic motivation, lower levels of boredom and better performance. In turn, sorting by image provides more autonomy and performance compared to the random approach. While we do not find a strictly dominant structure among the options of sorting by image or by label, we can however conclude, that either dominates the random sorting approach. This allows interactive labeling system designers to evaluate the present trade-offs and decide on the basis that fits their specific goals. It matters however, that they choose to provide such structure.

7.5 Limitations and Future Work

Our research has some limitations that we attempt to reduce in the course of the study. In particular, some of the constructs measured in the online experiment are complex. Given this, we relied only on beforehand validated measurement scales. To mitigate the risk of inattentive participants, exploitative behavior or other threats (e.g. the risk of fully automated bots answering the survey), we included attention check questions and applied additional measures to ensure data quality (see Section 6.1). In addition, the online experiment was performed in a single session. While this study allowed us to comprehend perceptions and performance behaviors when using the proposed labeling system, it also opened the space to question whether these remain stable when the system is used more than once by the same group of crowd workers. In particular, we see the potential for studying longer-term benefits. Therefore, we suggest that future research replicate our study but allow crowd workers to use the system more than once on different days and collect data over a series of days.

Hereby, we performed the online experiment with a fully functional labeling artifact. While we were able to show that structured approaches dominate over random sorting in crowdsourced labeling tasks in terms of task variety, autonomy, intrinsic motivation, boredom and performance, we were unable to identify a clear dominant approach. Therefore, in the future, we propose to compare sorting by image and sorting by label in further experimental contexts and task designs. In addition, other features such as gamification elements (e.g. points, badges and rank positions) could be of interest in the context of labeling (Knaeble *et al.*, 2020). Thus, we recommend that future research examine such elements on top of those examined in this study.

Finally, although we followed the established assumptions of SDT, we are unable to find an effect on performance in our path analysis (cf. Figure 6). Therefore, we believe that labeling performance depends on a third variable not considered in our study, which needs to be uncovered in future work.

8 Conclusion

Where research on ML algorithms makes steady progress, and approaches like deep learning call for more training data than ever before, researchers often disregard the origin of such data. Especially how their labels come to be attached to images, videos or other types of data. More often than not, this is done by manual labour in heteromated settings like crowd work (Ekbia & Nardi, 2014). With our research, we want to highlight the effort of labeling workers going towards the creation of future artificial intelligence. We present a new, but highly relevant area of research for SDT. Labeling sets out to become a new blue-collar job (Knaeble *et al.*, 2023, Reese, 2016) in crowdsourced micro gigs from homes all over the world. Thus, research must understand the specifics of such tasks—and SDT offers a theoretical lens to do so. With our work we further bridge existing SDT-based knowledge to mandatory use contexts in labeling and crowd work. Thereby, we specifically focus on labeling workers on crowdsourcing platforms. We show how merely changing the presentation sequence of labeling tasks provides benefits and identify a surprising contrast between mathematical and perceived task variety. Paying increased attention to labeling workers, in both academia and industry, could lead to not only better work results, but also help motivating the many workers and increase their well-being. To recapitulate on our visual metaphor: if you assign labeling tasks, consider showing your workers paintings, not noise.

Funding

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number 447287107).

References

Amabile, T. M. (1996) *Creativity in Context: Update to "The Social Psychology of Creativity"*, p. xviii, 317. Westview Press.

Amershi, S., Cakmak, M., Knox, W. B. and Kulesza, T. (2014) Power to the people: the role of humans in interactive machine learning. *AI Mag.*, **35**, 105. https://doi.org/10.1609/aimag.v35i4.2513.

Asdecker, B. and Zirkelbach, F. (2020) What drives the drivers? A qualitative perspective on what motivates the crowd delivery workforce. *Proceedings of the 53th Annual Hawaii International Conference on System Sciences (HICSS)*, 4011–4020.

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T. and Graesser, A. C. (2010) Better to be frustrated than bored: the incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *Int. J. Hum. Comput. Stud.*, **68**, 223–241.

Ballou, N., Deterding, S., Tyack, A., Mekler, E. D., Calvo, R. A., Peters, D., Villalobos-Zúñiga, G. and Turkay, S. (2022). Self-determination theory in HCI: shaping a research agenda. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3491101.3503702.

Bar-Hillel, M. and Wagenaar, W. A. (1991) The perception of randomness. *Adv. Appl. Math.*, **12**, 428–454. https://doi.org/10.1016/0196-8858(91)90029-I.

Benke, I., Gnewuch, U. and Maedche, A. (2022) Understanding the impact of control levels over emotion-aware chatbots. *Comput. Hum. Behav.*, **129**, 107122. https://doi.org/10.1016/j.chb.2021.107122.

Bentler, P. M. (1972) A lower-bound method for the dimension-free measurement of internal consistency. *Soc. Sci. Res.*, **1**, 343–357. https://doi.org/10.1016/0049-089X(72)90082-8.

Bentler, P. M. (2008) Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, **74**, 137–143. https://doi.org/10.1007/s11336-008-9100-1.

Bernard, J., Hutter, M., Zeppelzauer, M., Fellner, D. and Sedlmair, M. (2018a) Comparing visual-interactive labeling with active learning: an experimental study. *IEEE Trans. Vis. Comput. Graph.*, **24**, 298–308. https://doi.org/10.1109/TVCG.2017.2744818.

Bernard, J., Zeppelzauer, M., Sedlmair, M. and Aigner, W. (2018b) VIAL: a unified process for visual interactive labeling. *Vis. Comput.*, **34**, 1189–1207. https://doi.org/10.1007/s00371-018-1500-3.

Bollen, K. A. (1980) Issues in the comparative measurement of political democracy. *Am. Sociol. Rev.*, **45**, 370–390. https://doi.org/10.2307/2095172.

Cakmak, M., Chao, C. and Thomaz, A. L. (2010) Designing interactions for robot active learners. *IEEE Trans. Auton. Ment. Dev.*, **2**, 108–118. https://doi.org/10.1109/TAMD.2010.2051030.

Chandy, R. K., Tellis, G. J., Macinnis, D. J. and Thaivanich, P. (2001) What to say when: advertising appeals in evolving markets. *Journal of Marketing Research (JMR)*, **38**, 399–414. https://doi.org/10.1509/jmkr.38.4.399.18908.

Chen, B.*et al.* (2015) Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motiv. Emot.*, **39**, 216–236. https://doi.org/10.1007/s11031-014-9450-1.

Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J. and Aragon, C. R. (2018) Using machine learning to support qualitative coding in social science: shifting the focus to ambiguity. *ACM Trans. Interact. Intell. Syst.*, **8**, 1–20. https://doi.org/10.1145/3185515.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). L. Erlbaum Associates.

Cohen, B. (2020) *The Hot Hand: The Mystery and Science of Streaks* (1st edn). Custom House.

Coll, C., Bier, R., Li, Z., Langenheder, S., Gorokhova, E. and Sobek, A. (2020) Association between aquatic micropollutant dissipation and river sediment bacterial communities. *Environ. Sci. Technol.*, **54**, 14380–14392. https://doi.org/10.1021/acs.est.0c04393.

Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334. https://doi.org/10.1007/BF02310555.

De Vreede, T., Raghavan, M. and De Vreede, G.-J. (2021) *Design Foundations for AI Assisted Decision Making: A Self Determination Theory Approach*. https://doi.org/10.24251/HICSS.2021.019.

Deci, E. L. and Ryan, R. M. (1985) *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer Science+Business Media. https://doi.org/10.1007/978-1-4899-2271-7.

Deci, E. L. and Ryan, R. M. (2000) The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. *Psychol. Inq.*, **11**, 227–268. https://doi.org/10.1207/S15327965PLI110401.

Deci, E. L. and Ryan, R. M. (2008) Self-determination theory: a macrotheory of human motivation, development, and health. *Can. Psychol.*, **49**, 182–185.

Deci, E. L., Nezlek, J. and Sheinman, L. (1981) Characteristics of the rewarder and intrinsic motivation of the rewardee. *J. Pers. Soc. Psychol.*, **40**, 1–10. https://doi.org/10.1037/0022-3514.40.1.1.

Deci, E. L., Olafsen, A. H. and Ryan, R. M. (2017) Self-determination theory in work organizations: the state of a science. *Annu. Rev. Organ. Psychol. Organ. Behav.*, **4**, 19–43. https://doi.org/10.1146/annurev-orgpsych-032516-113108.

Derfler-Rozin, R., Moore, C. and Staats, B. R. (2016) Reducing organizational rule breaking through task variety: how task design supports deliberative thinking. *Organ. Sci.*, **27**, 1361–1379. https://doi.org/10.1287/orsc.2016.1094.

Deterding, S. (2016). Contextual autonomy support in video game play: a grounded theory. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3931–3943. https://doi.org/10.1145/2858036.2858395

Dimmock, J., Jackson, B., Podlog, L. and Magaraggia, C. (2013) The effect of variety expectations on interest, enjoyment, and locus of causality in exercise. *Motiv. Emot.*, **37**, 146–153. https://doi.org/10.1007/s11031-012-9294-5.

Dobler, D., Friedrich, S. and Pauly, M. (2018) Nonparametric MANOVA in Mann–Whitney effects. https://arxiv.org/abs/1712.06983.

Dobler, D., Friedrich, S. and Pauly, M. (2020) Nonparametric MANOVA in meaningful effects. *Ann. Inst. Stat. Math.*, **72**, 997–1022. https://doi.org/10.1007/s10463-019-00717-.

Dudley, J. J. and Kristensson, P. O. (2018) A review of user interface design for interactive machine learning. *ACM Trans. Interact. Intell. Syst.*, **8**, 1–37. https://doi.org/10.1145/3185517.

Durward, D., Blohm, I. and Leimeister, J. M. (2020) The nature of crowd work and its effects on individuals' work perception. *J. Manag. Inf. Syst.*, **37**, 66–95. https://doi.org/10.1080/07421222.2019.1705506.

Ekbia, H. and Nardi, B. (2014) Heteromation and its (dis)contents: the invisible division of labor between humans and machines. *First Monday*. https://doi.org/10.5210/fm.v19i6.5331.

Faul, F., Erdfelder, E., Lang, A.-G. and Buchner, A. (2007) G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods*, **39**, 175–191. https://doi.org/10.3758/BF03193146.

Faul, F., Erdfelder, E., Buchner, A. and Lang, A.-G. (2009) Statistical power analyses using g*power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods*, **41**, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Findlater, L. and McGrenere, J. (2010) Beyond performance: feature awareness in personalized interfaces. *Int. J. Hum. Comput. Stud.*, **68**, 121–137. https://doi.org/10.1016/j.ijhcs.2009.10.002.

Flink, C., Boggiano, A. K. and Barrett, M. (1990) Controlling teaching strategies: undermining children's self-determination and performance. *J. Pers. Soc. Psychol.*, **59**, 916–924. https://doi.org/10.1037/0022-3514.59.5.916.

Fornell, C. and Larcker, D. F. (1981) Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.*, **18**, 39–50. https://doi.org/10.2307/3151312.

Fujimoto, Y., Ferdous, A. S., Sekiguchi, T. and Sugianto, L.-F. (2016) The effect of mobile technology usage on work engagement and emotional exhaustion in Japan. *J. Bus. Res.*, **69**, 3315–3323. https://doi.org/10.1016/j.jbusres.2016.02.013.

Gagné, M., Parent-Rocheleau, X., Bujold, A., Gaudet, M.-C. and Lirio, P. (2022a) How algorithmic management influences worker motivation: a self-determination theory perspective. *Can. Psychol.*, **63**, 247–260. https://doi.org/10.1037/cap0000324.

Gagné, M., Parker, S. K., Griffin, M. A., Dunlop, P. D., Knight, C., Klonek, F. E. and Parent-Rocheleau, X. (2022b) Understanding and shaping the future of work with self-determination theory. *Nat. Rev. Psychol.*, **1**, 378–392. https://doi.org/10.1038/s44159-022-00056-w.

Gneezy, U. and Rustichini, A. (2000) Pay enough or don't pay at all. *Q. J. Econ.*, **115**, 791–810.

Guay, F., Vallerand, R. and Blanchard, C. (2000) On the assessment of situational intrinsic and extrinsic motivation: the situational motivation scale (SIMS). *Motiv. Emot.*, **24**, 175–213. https://doi.org/10.1023/A:1005614228250.

Hackman, J. and Oldham, G. R. (1976) Motivation through the design of work: test of a theory. *Organ. Behav. Hum. Perf.*, **16**, 250–279. https://doi.org/10.1016/0030-5073(76)90016-7.

Hahn, U. and Warren, P. A. (2009) Perceptions of randomness: why three heads are better than four. *Psychol. Rev.*, **116**, 454–461. https://doi.org/10.1037/a0015241.

Hair, J. F. (2019) *Multivariate Data Analysis* (8th edn). Cengage.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C. and Bigham, J. P. (2018) A data-driven analysis of workers' earnings on amazon mechanical turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems—CHI '18*, 1–14. https://doi.org/10.1145/3173574.3174023

Hassenzahl, M. (2008) User experience (UX): Towards an experiential perspective on product quality. *Proceedings of the 20th Conference on l'Interaction Homme-Machine*, 11–15. https://doi.org/10.1145/1512714.1512717

Haug, S., Benke, I. and Maedche, A. (2023) Aligning crowdworker perspectives and feedback outcomes in crowd-feedback system design. *Proc. ACM Hum.-Comput. Interact.*, **7**, 1–28. https://doi.org/10.1145/3579456.

Hettiachchi, D., van Berkel, N., Kostakos, V. and Goncalves, J. (2020) CrowdCog: a cognitive skill based system for heterogeneous task assignment and recommendation in crowdsourcing. *Proc. ACM Hum.-Comput. Interact.*, **4**, 1–22. https://doi.org/10.1145/3415181.

van Hooff, M. L. M. and van Hooft, E. A. J. (2017) Boredom at work: towards a dynamic spillover model of need satisfaction, work motivation, and work-related boredom. *Eur. J. Work Organ. Psychol.*, **26**, 133–148. https://doi.org/10.1080/1359432X.2016.1241769.

Huiberts, S. (2019) Every pixel has a random color. https://commons.wikimedia.org/wiki/File:Everypixelhasarandomcolor.png.

Javalgi, R. G., Cutler, B. D. and Malhotra, N. K. (1995) Print advertising at the component level: a cross-cultural comparison of the United States and Japan. *J. Bus. Res.*, **34**, 117–124. https://doi.org/10.1016/0148-2963(94)00116-V.

Jing, J., dr Angremont, E., Zafar, S., Rosenthal, E. S., Tabaeizadeh, M., Ebrahim, S., Dauwels, J. and Westover, M. B. (2018) Rapid annotation of seizures and interictal-ictal continuum EEG patterns. *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3394–3397. https://doi.org/10.1109/EMBC.2018.8513059

Knaeble, M., Nadj, M. and Maedche, A. (2020) Oracle or teacher? A systematic overview of research on interactive labeling for machine learning. *Proceedings of the 15th International Conference on Wirtschaftsinformatik (WI2020, Zentrale Tracks)*, 2–16. https://doi.org/10.30844/wi2020a1-knaeble

Knaeble, M., Nadj, M., Germann, L. and Maedche, A. (2023) Tools of trade of the next blue-collar job? antecedents, design features, and outcomes of interactive labeling systems. *31st European Conference on Information Systems (ECIS) Research Papers*. https://aisel.aisnet.org/ecis2023rp/373

Kohn, A. (1993) *Punished by Rewards: The Trouble With Gold Stars, Incentive Plans, a's, Praise, and Other Bribes*. Company, Houghton, Mifflin.

Krening, S. and Feigh, K. M. (2019) Effect of interaction design on the human experience with interactive reinforcement learning. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 1089–1100. https://doi.org/10.1145/3322276.3322379

Lamprinou, D. and Paraskeva, F. (2015) Gamification design framework based on SDT for student motivation. *2015 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL)*, 406–410. https://doi.org/10.1109/IMCTL.2015.7359631

Lee, D. (2018) *Why Big Tech Pays Poor Kenyans to Teach Self-Driving Cars*. BBC News. Retrieved April 25, 2022, fromhttps://www.bbc.com/news/technology-46055595.

Leimeister, J. M., Huber, M., Bretschneider, U. and Krcmar, H. (2009) Leveraging crowdsourcing: activation-supporting components for IT-based ideas competition. *J. Manag. Inf. Syst.*, **26**, 197–224. https://doi.org/10.2753/MIS0742-1222260108.

Leong, T. W., Vetere, F. and Howard, S. (2006). Randomness as a resource for design. *Proceedings of the 6th Conference on Designing Interactive Systems*, 132–139. https://doi.org/10.1145/1142405.1142428.

Liang, H., Peng, Z., Xue, Y., Guo, X. and Wang, N. (2015) Employees' exploration of complex systems: an integrative view. *J. Manag. Inf. Syst.*, **32**, 322–357. https://doi.org/10.1080/07421222.2015.1029402.

Liang, H., Wang, M.-M., Wang, J.-J. and Xue, Y. (2018) How intrinsic motivation and extrinsic incentives affect task effort in crowdsourcing contests: a mediated moderation model. *Comput. Hum. Behav.*, **81**, 168–176. https://doi.org/10.1016/j.chb.2017.11.040.

Liu, Y., Guo, B., Wang, Y., Wu, W., Yu, Z. and Zhang, D. (2016) TaskMe: multi-task allocation in mobile crowd sensing. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 403–414. https://doi.org/10.1145/2971648.2971709

Ma, X., Khansa, L. and Kim, S. S. (2018) Active community participation and crowdworking turnover: a longitudinal model and empirical test of three mechanisms. *J. Manag. Inf. Syst.*, **35**, 1154–1187. https://doi.org/10.1080/07421222.2018.1523587.

Manganelli, L., Thibault-Landry, A., Forest, J. and Carpentier, J. (2018) Self-determination theory can help you generate performance and well-being in the workplace: a review of the literature. *Adv. Dev. Hum. Resour.*, **20**, 227–240. https://doi.org/10.1177/1523422318757210.

Marx, K. and Engels, F. (1932) *The German Ideology*. David Riazanov (Marx–Engels–Lenin Institute).

McDonald, R. P. (1999) *Test Theory: A Unified Treatment*. L. Erlbaum Associates. https://doi.org/10.4324/9781410601087.

Milyavskaya, M. and Koestner, R. (2011) Psychological needs, motivation, and well-being: a test of self-determination theory across multiple domains. *Personal. Individ. Differ.*, **50**, 387–391. https://doi.org/10.1016/j.paid.2010.10.029.

Mondriaan, P. (1921) *Compositie met groot rood vlak, geel, zwart, grijs en blauw*. Kunstmuseum Den Haag. https://www.kunstmuseum.nl/nl/collectie/compositie-met-groot-rood-vlak-geel-zwart-grijs-en-blauw.

Morgeson, F. P. and Humphrey, S. E. (2006) The work design questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. *J. Appl. Psychol.*, **91**, 1321–1339. https://doi.org/10.1037/0021-9010.91.6.1321.

Mouzas, S. (2006) Efficiency versus effectiveness in business networks. *J. Bus. Res.*, **59**, 1124–1132. https://doi.org/10.1016/j.jbusres.2006.09.018.

Nadj, M., Knaeble, M., Li, M. X. and Maedche, A. (2020) Power to the oracle? Design principles for interactive labeling systems in machine learning. *KI - Künstliche Intelligenz*, **34**, 131–142. https://doi.org/10.1007/s13218-020-00634-1.

Nickerson, R. S. (2002) The production and perception of randomness. *Psychol. Rev.*, **109**, 330–357. https://doi.org/10.1037/0033-295X.109.2.330.

Ntoumanis, N. and Standage, M. (2009) Motivation in physical education classes: a self-determination theory perspective. *Theory Res. Educ.*, **7**, 194–202. https://doi.org/10.1177/1477878509104324.

Orzech, K. M., Grandner, M. A., Roane, B. M. and Carskadon, M. A. (2016) Digital media use in the 2 h before bedtime is associated with sleep variables in university students. *Comput. Hum. Behav.*, **55**, 43–50. https://doi.org/10.1016/j.chb.2015.08.049.

Pentland, B. T. (2003) Conceptualizing and measuring variety in the execution of organizational work processes. *Manag. Sci.*, **49**, 857–870. https://doi.org/10.1287/mnsc.49.7.857.16382.

Peters, D., Calvo, R. A. and Ryan, R. M. (2018) Designing for motivation, engagement and wellbeing in digital experience. *Front. Psychol.*, **9**, 797. Retrieved September 6, 2022, fromhttps://www.frontiersin.org/articles/10.3389/fpsyg.2018.00797.

Qiao, D., Lee, S.-Y., Whinston, A. B. and Wei, Q. (2021) Mitigating the adverse effect of monetary incentives on voluntary contributions online. *J. Manag. Inf. Syst.*, **38**, 82–107. https://doi.org/10.1080/07421222.2021.1870385.

Reese, H. (2016) *Is 'Data Labeling' the New Blue-Collar Job of the AI Era?* TechRepublic. Retrieved July 12, 2021, fromhttps://www.techrepublic.com/article/is-data-labeling-the-new-blue-collar-job-of-the-ai-era/.

Renson, R. and Careel, C. (1986) Sporticuous consumption: an analysis of social status symbolism in sport ads. *Int. Rev. Sociol. Sport*, **21**, 153–171. https://doi.org/10.1177/101269028602100207.

Rosseel, Y. (2012) Lavaan: an r package for structural equation modeling. *J. Stat. Softw.*, **48**, 1–36. https://doi.org/10.18637/jss.v048.i02.

Ryan, R. M. and Deci, E. L. (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.*, **55**, 68–78. https://doi.org/10.1037/0003-066X.55.1.68.

Ryan, R. M. and Grolnick, W. (1986) Origins and pawns in the classroom: self-report and projective assessment of individual differences in children's perceptions. *J. Pers. Soc. Psychol.*, **50**, 550–558. https://doi.org/10.1037/0022-3514.50.3.550.

Saito, S., Chiang, C.-W., Savage, S., Nakano, T., Kobayashi, T. and Bigham, J. P. (2019) TurkScanner: predicting the hourly wage of microtasks. *The World Wide Web Conference—WWW '19*, 3187–3193. https://doi.org/10.1145/3308558.3313716.

Schultz, P. P., Ryan, R. M., Niemiec, C. P., Legate, N. and Williams, G. C. (2015) Mindfulness, work climate, and psychological need satisfaction in employee well-being. *Mindfulness*, **6**, 971–985. https://doi.org/10.1007/s12671-014-0338-7.

Settles, B. (2009) *Active Learning Literature Survey*. Computer Sciences Technical Report No. 1648. University of Wisconsin-Madison, Madison, WI.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sheller, M. (2004) Automotive emotions: feeling the car. *Theory Cult. Soc.*, **21**, 221–242. https://doi.org/10.1177/0263276404046068.

van der Stappen, A. and Funk, M. (2021) Towards guidelines for designing human-in-the-loop machine training interfaces. *26th International Conference on Intelligent User Interfaces*, 514–519. https://doi.org/10.1145/3397481.3450668

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J. and Weich, S. (2009) Internal construct validity of the Warwick–Edinburgh mental well-being scale (WEMWBS): a rasch analysis using data from the scottish health education population survey. *Health Qual. Life Outcomes*, **7**, 1–8. https://doi.org/10.1186/1477-7525-7-15.

Szalma, J. L. (2014) On the application of motivation theory to human factors/ergonomics: motivational design principles for human–technology interaction. *Hum. Factors*, **56**, 1453–1471. https://doi.org/10.1177/0018720814553471.

Tellis, G. J., MacInnis, D. J., Tirunillai, S. and Zhang, Y. (2019) What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence. *J. Mark.*, **83**, 1–20. https://doi.org/10.1177/0022242919841034.

Thomas, K. A. and Clifford, S. (2017) Validity and mechanical turk: an assessment of exclusion methods and interactive experiments. *Comput. Hum. Behav.*, **77**, 184–197. https://doi.org/10.1016/j.chb.2017.08.038.

Tian, Y., Liu, W., Xiao, R., Wen, F. and Tang, X. (2007) A face annotation framework with partial clustering and interactive labeling. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2007.383282.

Tuttle, T. C. and Romanowski, J. J. (1985) Assessing performance and productivity in white-collar organizations. *National Productivity Review*, **4**, 211–224. https://doi.org/10.1002/npr.4040040302.

Tyack, A. and Mekler, E. D. (2020) Self-determination theory in HCI games research: current uses and open questions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3313831.3376723

Van den Broeck, A., Vansteenkiste, M., De Witte, H., Soenens, B. and Lens, W. (2010) Capturing autonomy, competence, and relatedness at work: construction and initial validation of the work-related basic need satisfaction scale. *J. Occup. Organ. Psychol.*, **83**, 981–1002. https://doi.org/10.1348/096317909X481382.

Villalobos-Zúñiga, G. and Cherubini, M. (2020) Apps that motivate: a taxonomy of app features based on self-determination theory. *Int. J. Hum. Comput. Stud.*, **140**, 102449. https://doi.org/10.1016/j.ijhcs.2020.102449.

Xiao, L. and Ding, M. (2014) Just the faces: exploring the effects of facial features in print advertising. *Mark. Sci.*, **33**, 338–352. https://EconPapers.repec.org/RePEc:inm:ormksc:v:33:y:2014:i:3:p:338-352.

Yuan, L. (2018) *How Cheap Labor Drives China's A.I. Ambitions*. The New York Times. Retrieved April 25, 2022, fromhttps://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html.

Zhang, L., Tong, Y. and Ji, Q. (2008) Active image labeling and its application to facial action labeling. In D. Forsyth, P. Torr and A. Zisserman (eds), *Computer Vision—ECCV 2008*, pp. 706–719. Springer. https://doi.org/10.1007/978-3-540-88688-452.

Zhang, Y., Wang, Y., Zhang, H., Zhu, B., Chen, S. and Zhang, D. (2022). OneLabeler: a flexible system for building data labeling tools. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3491102.3517612

APPENDIX

Control Variables

Measurement

To account for extraneous variation, we account for standard demographics (age, gender, first language, and highest level of education), as well as previous experience. We have the participants enter their age, measured in years. Gender has to be selected from a choice of three (male, female, diverse). The highest level of education is selected from a choice of five (below high school equivalent, high school equivalent, bachelor's degree, master's degree, doctoral degree). The first language shows a choice of the most common languages as predefined options and a free-text alternative, if a language is not listed. Furthermore, we ask the participants to rate their previous experiences with the advertising industry, emotional cues, labeling tasks and ML on a seven-point semantic differential scale.

Moreover, we controlled for other structural aspects of the labeling task. Hereby, we randomly introduced half of our participants to slightly modified versions of the artifact. Firstly, we changed the button order. Whereas initially the left side of the selector represented yes/positive, and the right side no/negative, we switched this order for some. Although the neutral default should already negate priming or default effects, we further controlled for the influence of this order. Additionally, we changed the size of each labeling step. While in total, all participants labeled the same amount of data, for half we increased the number of labels assigned on one screen to four (from initially one). Hereby, another structural component of the labeling process is controlled for.

Table A1. Convergent validity indicators of average variance extracted (AVE), Ω 1 Bollen (1980), 2 Bentler (1972) and 3 McDonald (1999) and Cronbach's α Cronbach (1951)

Construct	AVE	Ω_1	Ω_2	Ω_3	Cr. α
Variety	0.862	0.962	0.962	0.955	0.963
Autonomy	0.650	0.879	0.879	0.889	0.868
Intrinsic m.	0.862	0.961	0.961	0.961	0.960
Well-being	0.709	0.923	0.923	0.925	0.918
Boredom	0.814	0.946	0.946	0.946	0.945

Table A2. Discriminant validity indicator correlations or heterotrait–monotrait ratio

	Va.	Au.	I. m.	W.-b.	Bo.
Variety	1				
Autonomy	0.298	1			
Intrinsic m.	0.377	0.620	1		
Well-being	0.062	0.168	0.338	1	
Boredom	0.313	0.539	0.608	0.249	1

Table A3. Insignificant direct effects

	p	β
T.V. → W.-b.	0.069	-0.119
Au. → W.-b.	0.720	-0.026
T.V. → Bd.	0.158	-0.083
T.V. → Pf.	0.137	-2.138
Au. → Pf.	0.983	-0.034
I.m. → Pf.	0.760	0.542

Hypothesis Testing

A higher experience with the advertisement industry has a significant positive effect on the task variety ($p = 0.037$, $\beta = 0.114$). Similarly, a higher experience with ML has a significant positive intrinsic motivation ($p = 0.047$, $\beta = 0.068$).

Higher age significantly affects variety ($p = 0.044$, $\beta = 0.004$) and autonomy ($p = 0.004$, $\beta = 0.005$) positively, as well as performance negatively ($p < 0.001$, $\beta = -0.117$).

Table A4. Items in questionnaire, all measured on a 7 pt. Likert scale. Indicator loadings for convergent validity analysis following the guidelines by Hair (2019). Variety based on Morgeson & Humphrey (2006), autonomy on Schultz *et al.* (2015), intrinsic motivation on Guay *et al.* (2000), well-being on Stewart-Brown *et al.* (2009) and boredom on van Hooff & van Hooft (2017).

Construct	Context-Adapted Phrasing	Loading
Variety	This activity involves a great deal of task variety	0.855
Variety	This activity involves doing a number of different things.	0.909
Variety	This activity requires the performance of a wide range of tasks.	0.969
Variety	This activity involves performing a variety of tasks.	0.980
Autonomy	In this activity, I feel a sense of choice and freedom in the things I undertake	0.613
Autonomy	I feel that my decisions in this activity reflect what I really want.	0.649
Autonomy	I feel my choices in this activity express who I really am.	0.630
Autonomy	I feel I have been doing what really interests me in this activity.	0.714
Autonomy	Most of the things I do in this activity feel like "I have to".	0.806
Autonomy	I feel forced to do many things in this activity I wouldnt choose to do	0.784
Autonomy	I feel pressured to do too many things in this activity.	0.572
Autonomy	This activity feels like a chain of obligations.	0.837
Intrinsic m.	I am engaged in this activity because I think that this activity is interesting	0.921
Intrinsic m.	I am engaged in this activity because I think that this activity is pleasant	0.940
Intrinsic m.	I am engaged in this activity because this activity is fun	0.946
Intrinsic m.	I am engaged in this activity because I feel good when doing this activity	0.899
Well-being	I've been feeling optimistic about the future during this activity.	0.804
Well-being	I've been feeling useful during this activity.	0.840
Well-being	I've been feeling relaxed during this activity.	0.883
Well-being	I've been dealing with problems well during this activity.	0.902
Well-being	I've been thinking clearly during this activity.	0.777
Well-being	I've been feeling close to other people during this activity.	0.694
Well-being	I've been able to make up my own mind about things during this activity.	0.668
Boredom	There are long periods of boredom during this activity.	0.924
Boredom	This activity went by slowly.	0.914
Boredom	I often got bored during this activity.	0.933
Boredom	The time seemed to go by slowly during this activity.	0.850

Identifying as male affects intrinsic motivation negatively ($p < 0.001$, $\beta = -0.051$), while a higher level of education affects boredom positively ($p = 0.034$, $\beta = 0.078$). Boredom is negatively affected by a larger step size ($p = 0.049$, $\beta = -0.053$), while inverting the yes/no button arrangement to a no/yes order negatively affects performance ($p = 0.033$, $\beta = -1.364$).

MANOVA

We find no significant differences between the groups with regards to our control variables. On our scale from 0 to 1, participants report their experiences on average as follows: advertisements 0.243, emotional cues 0.257, labeling 0.519 and ML 0.461.