Heterogeneous agent decision-making - an empirically informed approach to behavioural types

Nanda Wijermans[1,2][0000−0003−4636−315X], Eva Vriens[1,3][0000−0002−8824−0812], and Giulia Andrighetto[1,3][0000−0002−3896−1363]

[1] Institute For Future Studies, Holländargatan 13, 11136, Stockholm, Sweden
[2] Stockholm Resilience Centre, Stockholm University Albanovägen 28,10691 Stockholm, Sweden
[3] The Institute for Cognitive Sciences and Technologies, Italian National Research Council, Via S. Martino della Battaglia 44, 00185 Rome, Italy
nanda.wijermans@su.se

Abstract. How to model human-decision making reflects one of the biggest challenges as modellers of social systems. In this paper we present our approach to modelling different types of decision-making agents designed to engage in a collective risk social dilemma experiment. Like their human counterparts in a controlled behavioural experiment, they are confronted with a choice of contributing to avoid a disaster under different levels of risk and do not know whether the others will also contribute. To design these agents, we make use of data from a conditional choice task conducted during the behavioural experiment combined with theory. Rather than fitting or calibrating on empirical data, the empirics mainly informs a narrative for the formalisation of four types of agents with different sensitivities to risk and norms. The paper ends with describing how we validate this model by comparing the multiple outcome measures with the behavioural patterns in the behavioural experiments.

Keywords: agent decision-making· empirics-based model design· Agent-based modelling · Decision-types · Social Dilemmas · Social Norms· Behavioural experiments

1 Introduction

Modelling human decision-making is at the heart of our models. It is a core challenge any modeller to decide and justify the fit of the decision model with the target phenomenon [9]. The use of empirics in agent-based social simulation is more and more conventional, however stressed in its use for calibration and validations[3]. The way we go about using empirics is not often at the forefront of our papers or discussions. In particular, empirical data to design our models is often associated with calibration (to fix/t the values of variables in a (often default go-to) decision model (e.g. bounded rational deciding in the form of maximising of some utility) that has been chosen. Through calibration is fine, it

is not the only entry into make use of empirics to design models. Empirics for model design can also involve the selection of variables or processes, supporting the formalisation process in which the richness of the target phenomenon is filtered for the key aspects and processed [6]. In this paper we describe our approach to design our agents using empirical data obtained via questionnaire during a behavioural experiment (conditional choice task). The cluster analysis of the answers leads us to describe four decision-making types. These narratives or agent profiles are then used to formalise the agents. We seek to contribute by sharing our approach, discuss this with our peers at the conference and to engage and learn with others about (other) approaches for agent design using empirics.

2 Background

The ABM we focus on in this paper is closely connected to an online behavioural experiment called the collective risk social dilemma experiment [7, 8]. Firstly, the ABM targets the same decision situation as the human participants have and thus reflects the same experimental game as the behavioural experiment. Secondly, in the design of the agent-decision making, we make use of behavioural patterns observed in the conditional choice tasks conducted during these experiments.

2.1 Decision context in behavioural experiment

The collective risk social dilemma experiment is designed to study the dynamics of social norms under collective risk, whether social norms causally motivate behaviour, and how this affects the ability of groups to solve cooperation problems. Social norms are defined as informal behavioural rules that individuals follow conditionally on their believing that: (i) a sufficiently large number of people in their community conform to the rule (empirical expectations), and (ii) a sufficiently large number of people in their community think that people ought to conform to the rule (normative expectations) [1]. The collective risk social dilemma experiment is described in detail in [7]. We only summarise its main features.

Participants played one round per day for a period of 28 days. Randomly matched in groups of six people, the participants are confronted with a collective risk social dilemma [5]. If they cooperate, a possible disaster is averted. If they fail to, with some (known) risk probability the disaster occurs and they lose everything. Every day participants are matched in a new group and are confronted with the same dilemma. Each day, participants are receive a constant initial endowment (100 points). Participants can avoid the risk of collective loss by investing part of this endowment into a public project that is able to protect them from the loss only if a minimal threshold is reached (in the experiment the threshold is 300 points). If the threshold is reached, the disaster is averted with

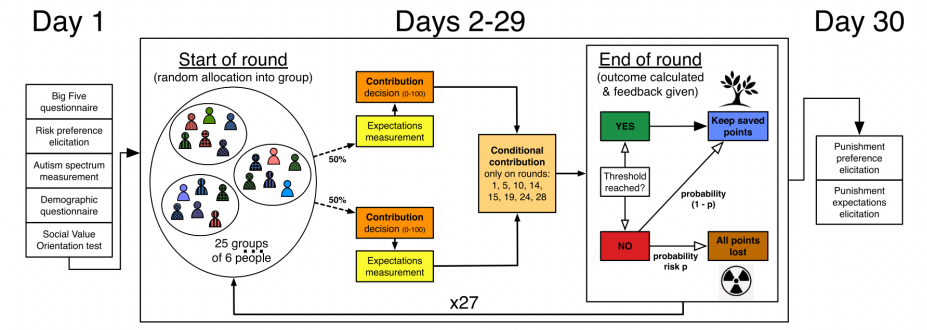

Fig. 1. Structure of the experimental set-up used in [7].

certainty and participants earn what they did not spent on preventing the disaster. If not, participants risk losing their earnings for that round with probability p. At the end of every round, participants are informed about the contribution of their group members, the outcome, whether a disaster occurred, and their individual payoff for that round.

To experiment aimed to test how risk affects social norms and cooperation, the risk probability (high: $p = 0.9$ vs low: $p = 0.6$) and the order in which participants face different risks were manipulated. Participants in treatment 1 (High-to-Low) experience high risk of disaster ($p = 0.9$) in rounds 1–14, followed by low risk ($p = 0.6$) in rounds 15–28. In Treatment 2 (Low-to-High) participants faced the two risks in the reversed order. Second, to diagnose the existence of social norms, empirical expectations and normative expectations were elicited in each of the 28 game-days to detect the basic conditions for norms to motivate behaviour [2]. Empirical expectations are participants' beliefs about what how many points others will contribute; while normative expectations are beliefs about how much others think that one ought to contribute. Finally, to identify the causal effects of empirical and normative expectations on behaviour, social expectations are manipulated in a "conditional contribution" phase (using the strategy method) in a subset of rounds. This asked participants how much they will contribute to the collective fund facing four combinations of high and low

Empirical Expectations and Normative Expectations: if the majority of their group members put in [at least 50 points $<$ 50 points] and believe that you should all spend [at least 50 points $<$ 50 points].

3 Empirics in model design: towards behavioural types.

The idea to develop an ABM was born when reflecting on the results of the collective risk social dilemma experiment by [7] and its replication conducted during the COVID-19 pandemic [8]. In both experiments, there was more cooperation and social norms were stronger when risk was higher ($p = 0.9$), and social norms were found to causally predict behaviour. However, the results also

show substantial heterogeneity, with participants responding differently to social norms. Moreover, groups that failed to reach the threshold (i.e., to contribute 300 tokens to the collective fund) not rarely they were only a few points short of reaching it, making cooperation very inefficient. This inspired us to study in more detail the individual-level decision-making strategies in response to norms and risk to design our ABM.

In our mission to advance the understanding of the role of norms under dynamic collective risk, we developed an ABM — norms@risk — to reflect the decision situation of the collective risk social dilemma experiment [7], as described in background (2). To explain the behaviour observed in the experiment, we designed agent types that are inspired by the diverse behavioural responses to the conditional choice tasks. Hence, rather than calibrating our ABM on the actual contribution choices made by the participants in the interactive experiment, we formalise strategies based on their responses in a different task and test whether this classification reproduces some of the behavioural dynamics observed in the experiments.

3.1 Empirical clusters

The main scope of the behavioural experiments was to understand how social norms and/or risk guide individual behaviour. In an interactive choice experiment, the endogenous evolution of social norms and individual behaviour were tracked for (a change in) exogenous imposed risks of disasters. On top of this interactive task, on 8 out of 28 days the participants were confronted with a conditional choice task that manipulated not only the exogenous risk level (0.6 or 0.9) but also the social norm in place. Four times under a 0.9 risk probability and four times under a 0.6 risk probability, their behaviour under a cooperative norm (non-cooperative norm) was solicited by asking how much they would contribute knowing that others would contribute more (less) than 50 points and think that one should contribute more (less) than 50 points.

We consider individual conditional contributions C^i_{xy} as the contribution of participant i, averaged across all conditional contribution answers, in the scenario in which the social norm is set to be x (h: > 50 points, l: < 50 points) and risk to be y (9: $p = 0.9$, 6: $p = 0.6$). The responsiveness to social norms (SNR) quantifies the change in contribution from a scenario in which the social norm is cooperative (h: > 50 points) compared to a scenario in which the social norm is non-cooperative (l: < 50 points) while keeping risk constant. The responsiveness to risk (RIR) compares the change in contribution from high (9: $p = 0.9$) to low risk (6: $p = 0.6$) while keeping the social norm fixed. Specifically, that means:

$$SNR_i = \frac{C^i_{h9} - C^i_{l9} + C^i_{h6} - C^i_{l6}}{2} \tag{1}$$

$$RIR_i = \frac{C^i_{h9} - C^i_{h6} + C^i_{l9} - C^i_{l6}}{2} \tag{2}$$

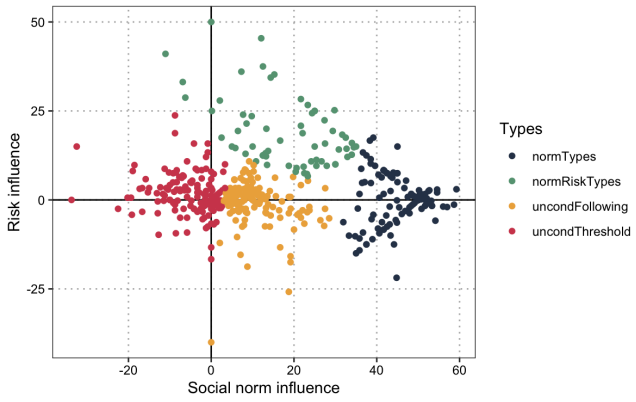

Fig. 2. Behavioural types mapped according to their responsiveness to social norms (SNR) and risk (RIR)

To get an insight on the sensitivity to norms and/or risk we conducted a cluster analysis on the combined data of [7] and [8]. Through K-means clustering, four meaningful clusters were extracted based on the individual scores on these two responsiveness dimensions (Figure 2). The majority of participants can be labelled unconditional cooperators (59%). They seem to cooperate out of intrinsic motivation and do not change their contribution much in relation to the risk level nor the social norm. With the difference that 28% cooperates unconditionally while slightly following the norm (increasing contribution by a few points when the norm is higher, decreasing when it is lower), while 31% is slightly affected by the norm in the opposite way. These 31% seem to be threshold driven as they increase the contribution compared to the social norm when the social norm is below the threshold, but decrease the contribution when the norm exceeds the threshold. The remaining 41% is affected by social norms and/or risk. 29% increases contributions when norms are high (social norm followers) and 12% responds to both norms and risk, contributing most when both are high.

Overall, we see the two types of unconditional participants always contributing around 50 points (both in the conditional contribution questions and during the real experiment), while the cluster of participants that respond to norms and/or risk lowers their contributions below 50 points when one or both of these conditions are not in place (see Figure 3). In the experiment, all participants contributed on average 50 points when risk was $p = 0.9$. The norm and norm&risk types slightly decreased contribution when risk is was $p = 0.6$, but not by much since norms were still moderately strong. Finally, a small minority (2.5%) of the types classified as responding to norm and/or risk contributed close to 0 when risk was $p = 0.6$—i.e., they became defectors. This suggests that as long as norms are present, cooperation is possible even with intermediate risk levels,

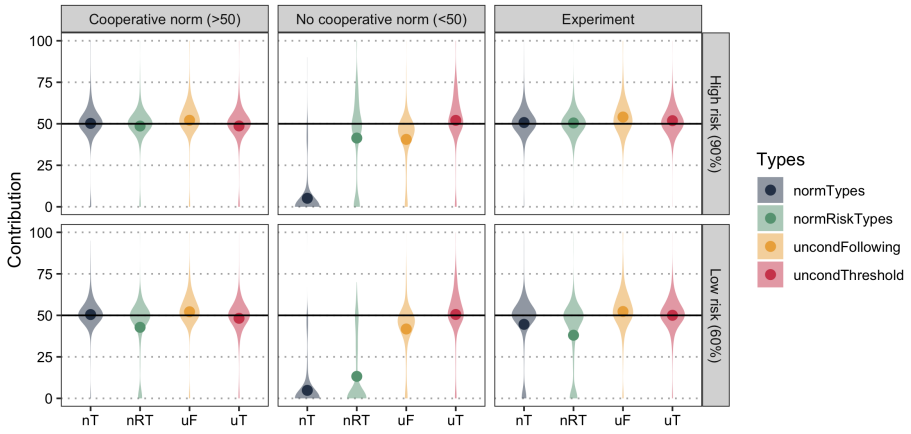

Fig. 3. Conditional and unconditional contributions of the four behavioural types for high and low risk and cooperative, non-cooperative, and real (unconditional) social norms.

but they do have to dictate to compensate for the lower contribution of the risk followers and defectors.

Type	Norm Sens	Risk sense	Behaviour
uncondFollowing	slightly	no	Contribute around 50 and when deviating in the direction of the norm
uncondThreshold	no	slightly	Contribute around 50 and when deviating in the direction of what is needed (less if others contribute more than enough, and more when there is a gap in what they think others will contribute
normTypes	yes	0	behaviour follows the norm
normRiskTypes	yes	yes	the type the is following the risk the strongest

Table 1. Table provides an overview of the different behavioural types that are reflected in the model.

3.2 From clusters to agents

The empirical behavioural clusters from form the basis for the formalisation of the different agent types: *unconditionalFollowers*, *unconditionalThresholders*, *normTypes* and *normRiskTypes*. The *unconditionals* (followers and thresholders) are types that want to reach the threshold and contribute around the fair share (personal tendency = fair share). Among the unconditionals, we distinguished

those that contribute slightly more/less following the norm (*uncondNfollowing type* or by contributing towards what is needed to reach the threshold given the norm, *uncondThreshold type*. The remaining agents are either *normTypes*, responding strongly to norms but relatively insensitive risk, or *normRiskTypes*, responding strongly to high levels of risk and nothing else. Note that following the empirical clusters a final agent type would be responsive to both norms and risk, but that is future work.

These agent types, in deciding whether and how much to contribute, they vary in their sensitivity to three aspects: 1) a personal tendency, 2) the current social norm and 3) the risk level. Resulting in the following formalisation of the contribution for each agent:

$$Contribution = w0 \cdot personalStrat + w1 \cdot normStrat + w2 \cdot riskStrat \quad (3)$$

The sum of the weights (w0, w1,w2) is normalised to equal 1. The strategies are described by:

$$pStrat = \{fairshare \pm xn, 0\}$$
$$nStrat = mean(ee, ne)$$
$$rStrat = riskPerception * endowment$$

The fair share reflects the threshold amount divided by the number of people in the group. The empirical expectations (ee) reflect agents' expectations of what others do based on what others have been contributing in the past. The normative expectations (ne) reflect agents' expectations of what others think one is supposed to contribute. The empirical and normative expectations are updated based on agents' experiences. The personal strategy remains rather stable, although the unconditionals adjust it by deviating slightly from xn by either following the norm or the threshold. To create the different types, the weights of the different strategies are set as follows:

- The defector is only sensitive to its personal strategy ($pStrat$, which is to contribute 0.
- The *unconditionals* are sensitive to the personal strategy that is reflecting their tendency to cooperate (= contribute the fair share), but are also in smaller extent sensitive to following the norm if they are *uncondNorm* or to towards the threshold *uncondThreshold*.
- The normTypes are only sensitive to the norm-strategy ($nStrat$)
- The riskTypes are only sensitive to the risk-strategy ($rStrat$)

We will now go a bit deeper into the perception of risk and norms.

3.3 Risk perception

Risk perception reflects an attribute level heterogeneity among the agents in how they perceive risk. Following the empirical behavioural clusters in which a minority of normRiskTypes decreased contribution for a risk of $p = 0.6$, we

initialised the agents with a distribution in which most view the risk as it is (subjective = objective risk), but some over- or underestimate it. Following the behavioural clusters, we initialised the normRiskTypes agents to underestimate risk more strongly than the rest of the population.

3.4 Norm perception

Norms in this model appear in three forms for each agent: personal norm (pn), empirical expectation (ee), and normative expectation (ne, following the formalisation of norms in the mathematical model of Gavrilets [4]. Personal norms reflect what is perceived the most appropriate contribution in a social setting, informing the personal strategy in this model. Empirical and normative expectations reflect the expectation about what others are contributing or the expectation what others think one should contribute. Each norm is updated over time with the same norm updating rate nur (see equation 4):

$$pn+ = nur * (contrib - pn) + (nur * (avg\_contrib\_others - pn)) \quad (4)$$

$$ee+ = nur * (ne - ee) + (nur * (avg\_contrib\_others - ee)) \quad (5)$$

$$ne+ = nur * (pn - ne) + (nur * (avg\_contrib\_others - ne)) \quad (6)$$

Note that future model extensions could vary the updating rates of different elements of the norm to make certain elements more important than others.

4 Do they behave? - Agent decision types' behaviour

To test whether the Norms@Risk model is useful (=validated) is when it can reproduce behavioural patterns in the behavioural experiment. To design a test of good-ness, we use a pattern-oriented approach - meaning that we use multiple patterns to test our model. We decided to define our patterns by using three outcome variables: the threshold percentage, the average individual contributions and the contribution expectations (the norm). This way we thus test our models on *group level* by comparing how groups overall perform,i.e. how many reach the threshold; on the *individual level* in terms of what are the contributions made on average by individuals and on the *cognitive level* how the expectations of contributions compare between the agents and the human experiment participants. At the same time we want to reflect on our models goodness in its overall performance as well as over time.

In the experiments the behavioural patterns can be formulated as follows:

Group level — threshold : The amount of groups reaching the threshold is higher when risk is higher (overall pattern G1); However within one risk level, the threshold percentage gets lower with low risk (0.6) in the high risk (0.9) conditions no substantial change is observed (overtime pattern G2).

Individual level — contribution : The higher the risk the higher the average contribution (overall - pattern I1). The contributions decrease over

time (within one risk level), in phase 1 of the game there seems to be a steeper decrease than in part 2, especially when the risk scenario is going from low-to-high (overtime pattern I2).

Cognitive level — norm Overall the expectations move with the risk level, like the contributions. in high risk (0.9) the expectations lie above 50 (meeting the threshold) and vice versa (overall pattern C1). Overtime the patterns are similar to the contributions, expectations goes down over time, especially in the first phase of the experiment (overtime pattern C2).

4.1 overall behaviour

For our validation test of the norms@risk model we ran an simulation experiment that mimics the empirical composition (what proportion of agent of each type) under different risk changes (*low-to-high* versus *high-to-low* risk scenario).

Starting with the overall picture showing the group level overview (outcome variables representing the overall run) in 4.

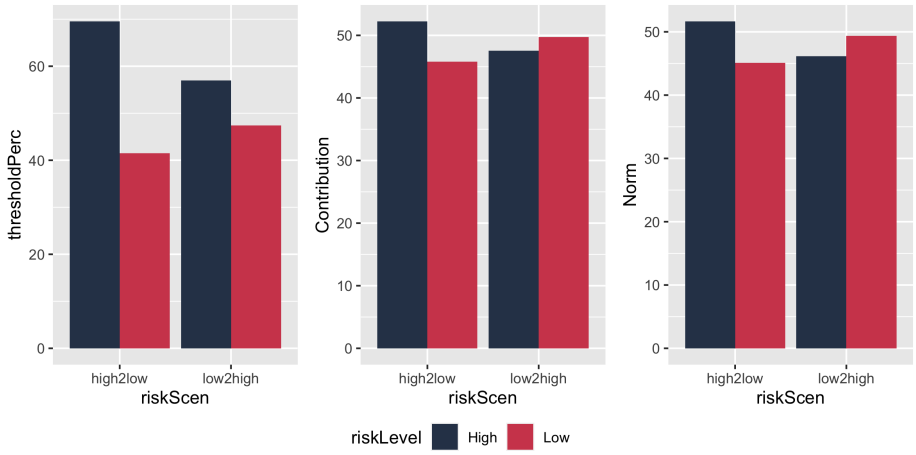

Fig. 4. Model behaviour on group level, how many groups made it to the threshold overall. Reflects the average of 1000 simulation repetitions

What we can see is that the threshold percentage, is indeed higher when risk is higher (overall pattern G1), and this difference is more pronounced in the *high2low* risk scenario. However, for the contribution this only holds in the high2low risk scenario, which is reflected in the norms/expectations too (overall - pattern I1).

4.2 Over time behaviour

To see what happens more closely and checking the overtime patterns we inspect figure 5 and 6

When we look at the threshold percentage over time (5), we see no pronounced change in threshold apart from a clear response to the risk level change in step 15 (dotted line). However within one risk level. When risk drops so does the percentage of groups making the threshold and vice versa. One sees an effect from what level the agents started with, but also from changing risk levels, but these patterns seem different especially going from low to high risk (overtime pattern G2).

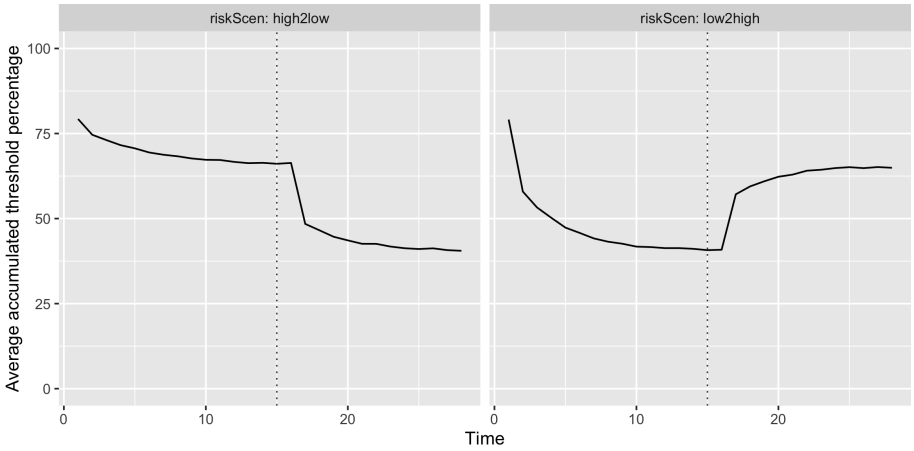

Fig. 5. The percentage of groups making it to the threshold each round. The vertical dotted line indicated the moment when the risk level changes. Reflects the average thresholdPerc of 1000 simulation repetitions

For the contributions over time (6) the decrease within one risk level is shown apart from going from low to high. Which also makes sense as they should be contributing more. The steeper decrease is not observable (overtime pattern I2). One reason for this could be that the behaviour shortly after the change of risk scenario should affect the agents more intensely than while being in a certain risk level for a while. Meaning that a mechanism would involve more than just responding to a different risk probability but also the fact that one is in a new situation and may be sensitivity to what others do as one accommodates for a new risk reality.

The reflections on what happens on the cognitive level - norm (7) are very similar to the contributions (overtime pattern C2). This is not surprising as the norms (expectations) are shaped after what is observed what others contribute.

Who contributes? Lastly we wanted to test on type-level how this compares to empirics. In particular, who contributes to when, what and how does this comprise the overall contribution level. We compare thus agents (figure 8) versus

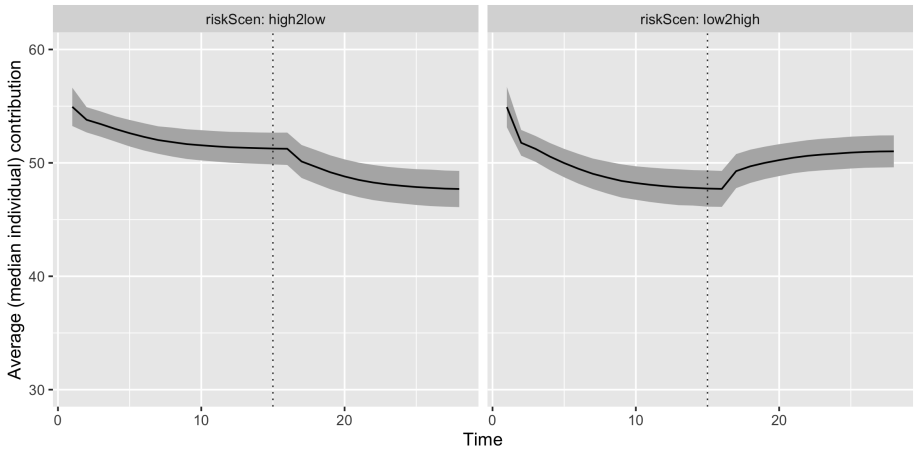

Fig. 6. Average contribution of individuals each round. The vertical dotted line indicated the moment when the risk level changes. Reflects the average individual contributions of 1000 simulation repetitions

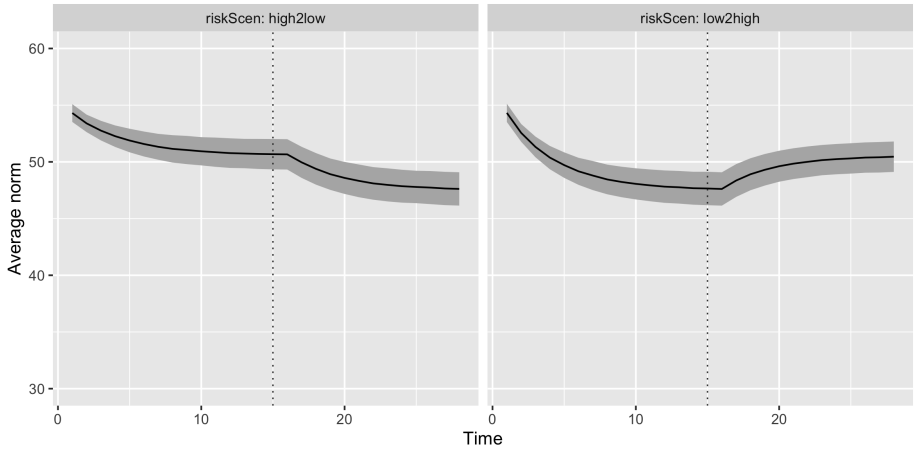

Fig. 7. The average norm (combination of expectation what others will do and what you think one should be doing) over time. The vertical dotted line indicated the moment when the risk level changes. Reflects the average norm of 1000 simulation repetitions

the empirical types (figure 9). What one can see is that the normRisk types (green line in both graphs) are the ones that push the change in overall behaviour in response to the risk change. The fact that these normRisk types do this is not a validation but a verification test, as this is how the decision type is designed. However the way the others follow is an emergent effect. For one, the range of the contributions by the types is more narrow under high risk than it is under low risk.

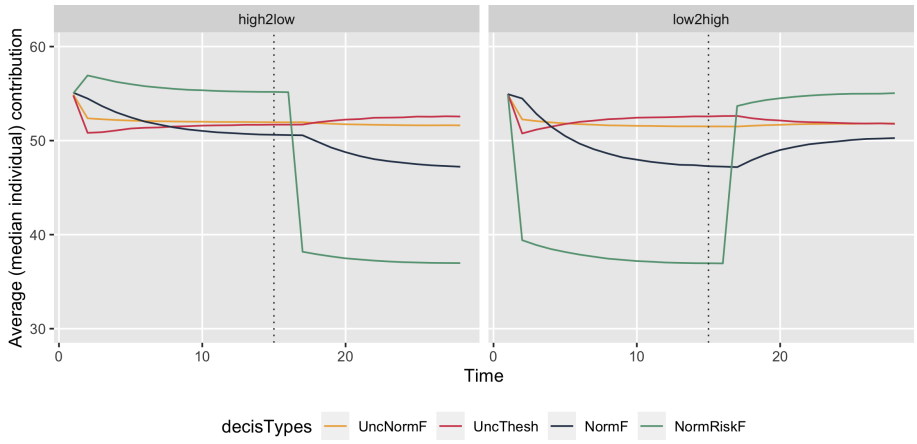

Fig. 8. Unpacks the average contribution of individuals per round per type. Reflects the average norm of 1000 simulation repetitions

4.3 Well behaved?

As the minimum, the model needed to able to 1) contribute more under high risk (and thus having more groups that make the threshold) than under lower risk; and 2) reproduce a response to the risk change affected by what happened in the past but adaptive to the new risk situation.

On this basic criterion the model passes and is good enough to explore other group compositions and risk levels to explore behaviours in settings that are impossible to do with an experiment (e.g. in the simulation we can play with the type compositions, in reality we cannot). However, there is a reason for having also multiple patterns is to reflect on mechanism validation from different angles. We strive for more from our agents in reflecting the decision making mechanisms. The validation results also show there is plenty of room for improvement. Especially the effect of risk change, versus the changes that happen when in a certain risk level are intriguing to dig deeper. These situational aspects of arriving in a new risk situation versus a being in risk situation could explain the different change in response. This concretely would mean a reconsideration of how

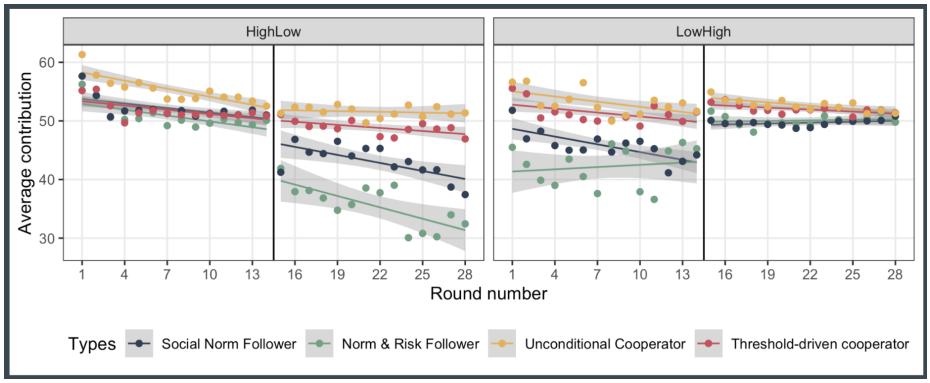

Fig. 9. The average contribution of individuals per round per type. Empirical analysis of behavioural data after identifying participants reflecting a certain type.

the updating of the norms work, or even more by introducing dynamics in the importance of what information is valued, e.g. making all agents for instance temporarily (more) norm-sensitive. Or even letting go of the whole type being a fixed characteristic but more a response to situation and past experience what type of decision maker is at a certain moment in time.

In short, do our agents behave? yes they do, but there is definitely room for improvement.

5 Conclusion

In this paper we presented the Norms@risk model, in particular how we designed the different decision-making types from empirical data and how we test its validity. Norms@risk is an agent-based model that targets to advance the understanding of the role of (changing) norms under dynamic risk. We detail the formalisation of empirical behavioural clusters into different agent decision types: unconditional types (uncondNorm and uncondThreshold types) that just want to contribute but are slightly sensitive to what others do, and those that are sensitive only to what others do (normTypes) and those sensitive to both others and the risk level (normRiskTypes). We then test how well the agents engage in a collective risk social dilemma, mimicking controlled behavioural experiments.

Discussion point for conference: how to determine validity? Recall that while validating our model we compared the empirical behavioural data with he simulation behavioural data. Our comparison is using *optical similarity*, by eye determining whether the graphs are deviating. We do this as the precise reproduction is not what we strife for, but more the general tendencies that get at the mechanisms. Or in other words, an exact match would almost be worrying and a difference of a certain number - it would be unclear what that even means. Also since the game setup is so sensitive around 50, the information whether an outcome variable is below or above 50 is more important than the numerical

nuances that cannot be spotted by a human eye. That said, there are many ways of comparing and welcome a discussion on the matter during the conference.

References

1. Bicchieri, C.: The grammar of society: The nature and dynamics of social norms. Cambridge University Press, Cambridge (2006)
2. Bicchieri, C., Lindemans, J.W., Jiang, T.: A structured approach to a diagnostic of collective practices. Frontiers in Psychology **5**(DEC), 1–13 (2014). https://doi.org/10.3389/fpsyg.2014.01418
3. Boero, R., Squazzoni, F.: Does Empirical Embeddedness Matter? Methodological Issues on Agent-Based Models for Analytical Social Science. jasss.soc.surrey.ac.uk pp. 1 – 31 (10 2005), http://jasss.soc.surrey.ac.uk/8/4/6.html
4. Gavrilets, S.: Coevolution of actions, personal norms and beliefs about others in social dilemmas. Evolutionary Human Sciences **3** (2021). https://doi.org/10.1017/ehs.2021.40
5. Milinski, M., Sommerfeld, R.D., Krambeck, H.J., Reed, F.A., Marotzke, J.: The collective-risk social dilemma and the prevention of simulated dangerous climate change. Proceedings of the National Academy of Sciences of the United States of America **105**(7), 2291–2294 (2008). https://doi.org/10.1073/pnas.0709546105
6. Schlüter, M., Wijermans, N., Elsler, L., González-Mon, B., Lindkvist, E., Martin, R., Martinez-Pena, R., Orach, K., Pellowe, K., Prawitz, H., Sanga, U.: Navigating the space between empirics and theory: empirical-stylized modelling of social-ecological systems. In: Proceedings of the Social Simulation Conference 2022
7. Szekely, A., Lipari, F., Antonioni, A., Paolucci, M., Sánchez, A., Tummolini, L., Andrighetto, G.: Evidence from a long-term experiment that collective risks change social norms and promote cooperation. Nature Communications **12**(1), 5452 (2021). https://doi.org/10.1038/s41467-021-25734-w
8. Vriens, E., Szekely, A., Lipary, F., Antonioni, A., Sánchez, A., Tummolini, L., Andrighetto, G.: Norms and cooperation under collective risk: A before-after comparison of Covid-19 pandemic threat in a long-term experiment. Working paper (2023)
9. Wijermans, N., Scholz, G., Chappin, E., Heppenstall, A., Filatova, T., Polhill, G., Semiuk, C., Stöppler, F.: Agent decision-making: the Elephant in the Room. Enabling justification of decision model fit in social-environmental models. Environmental Modelling and Software (under review)