An Experimental Attempt at Validating an Agent-Based Model on Decision-Making, Social Norm Change, and Norm Internalization

Marlene Batzke [0000-0001-5882-9813] and Andreas Ernst [0000-0001-5773-4441]

Center for Environmental Systems Research, University of Kassel, 34109 Kassel, Germany

Abstract. Understanding norm internalization remains one of the key questions that are still open in norm research. After formalizing a theory on norm internalization, implementing it into an agent-based model, and conducting an experimental study on norm internalization, the present work attempts at comparing experimental and simulation data. The agent-based DINO model simulates agents' decision-making and actions, social norms, and norm internalization in a 3-person Prisoners' Dilemma Game. The experiment was designed to match the model data. *N* = 365 participants were invited to play the structurally same game, while their social and personal norms were assessed repeatedly.

Participants' and agents' behavior, social norms, and norm internalization processes are compared regarding different social settings (cooperative vs. defective) and their willingness to cooperate (cooperator vs. conditional cooperator vs. defector). Results generally show substantial similarities between agents' and study participants' conditional cooperators, making the DINO model a valuable candidate for further testing and exploration. The comparison further suggested one mechanism in norm internalization that was so far missing and was therefore added to the DINO norm internalization process: asymmetry in internalizing cooperativeness verses defectivity. Further mechanisms in norm internalization and limitations of the comparison are discussed.

Keywords: Social Norms, Norm Internalization, Decision-Making, Learning, Social Dilemma, Cooperation.

1 Introduction

The question of how norms are internalized and internalized norms change is one of the key questions still open in norm research. Norm internalization describes the process of how individuals adopt and change their personal norms, being an individual's beliefs about the (in)appropriateness of a behavior in a specific situation. These personal norms can be differentiated from social norms, being beliefs about what other consider appropriate or normal behavior (Bicchieri et al., 2018; Cialdini et al., 1990). The power of social norms has long been known (Asch, 1956; Deutsch & Gerard, 1955; Sherif, 1936), influencing small group cooperation (Ostrom, 2000) and creating tipping points for large-scale transformations (Nyborg et al., 2016).

Yet, many authors have ascribed particular significance to norm internalization, making norm compliance independent from social norms and important for norm maintenance and long-term behavior change (Axelrod, 1986; Gintis, 2004). Studies have shown that the behavioral influence of social norms is largely mediated by personal norms (Hopper & Nielsen, 1991; Thøgersen, 1999). This may lead to the assumption that personal norms are influenced by social norms, yet particularly important for long-term behavior change (Otto & Kaiser, 2014) and decisions in the absence of social norm enforcement (Thøgersen, 2006). However, so far there are few theories that described the norm internalization process (Neumann, 2010), few simulation models that conceptualized it (Batzke & Ernst, 2023), and even fewer empirical studies that investigated it (Bamberg & Möser, 2007).

Norm internalization can be regarded as the product of the complex interplay of individuals' goals, habits, behaviors, et cetera, interacting with the social and physical environment over time, making internalization a suitable candidate to be studied via agent-based simulation. While simulation models on norm internalization may uniquely contribute to the understanding of the underlying mechanisms and dynamics (Andrighetto et al., 2010; Villatoro et al., 2015), it needs a combination of simulation and empirical data to further advance the study and understanding of norm internalization.

The present work provides a first attempt at comparing data on norm internalization from an agent-based model with experimental data. The conducted experiment was designed to produce data about variables matching those from the model. This allows partly testing, potentially validating, and improving the agent-based model. A psychologically grounded theory of decision-making, including social norms, goals, and habits was implemented in an agent-based model in the context of a social dilemma game. Hence, the present approach also allows comparing behavioral data and social norm change processes, contributing to the understanding of decision-making and social norm change in a social dilemma situation.

In the following, the implemented agent-based model and the conducted experimental study are presented. Then, model and experimental data are compared regarding participants' and agents' behavior, social norms, and personal norms. Finally, results are discussed.

2 An Agent-Based Model

The agent-based model DINO model (*D*ynamics of *I*nternalization and Dissemination of *No*rms) is presented and tested regarding the conditions and effects of norm internalization in Batzke and Ernst (2023). The model simulates the behavior of three agents in a 3-person *Prisoner's Dilemma Game* (see Dawes, 1980), describing the core of a conflict common to many situations. DINO agents' decision-making is determined by a weighted multi-attribute utility matrix, which represents goals, social norms, and habits as motivational factors in decision-making. There are three types of goals represented according to Deutsch (1958): the individualistic, cooperative, and competitive goal. Moreover, there are two types of social norms implemented, according to Cialdini et al. (1990): social descriptive norms (i.e., what others *do*) and social injunctive norms (i.e., what others (dis)approve).[1] All motivational factors are defined by a situational expectation and a personal value factor, along the *Theory of Planned Behavior* (Ajzen, 1991) and other expectation-value theories (Atkinson 1957; Fishbein & Ajzen, 1975). Situational factors are adapted over time, based on agents' experiences. Personal value factors represent the individual importance of a motivational factor and are represented statically (see Chapter 4). Agents' behavior is determined by their intention, defined as the weighted sum (i.e., expectation-value products) of all motivational factors.

Over and above the adaptation of situational expectations, which are assumed to be made rather quickly whenever the situation changes, the norm internalization process was implemented as a slow adaptation process, representing a more aggregated form of learning, depending on DINO agents' personal values and their experiences. Change in personal norms depends on agents' normative judgement regarding their last chosen action. Based on this evaluation, agents adapt their personal norms in a stepwise process. Personal norms influence decision-making through emphasizing or inhibiting the importance of the other motivational factors, representing a higher-level factor in decision-making (for details see Batzke & Ernst, 2023).

3 An Experiment with Participants

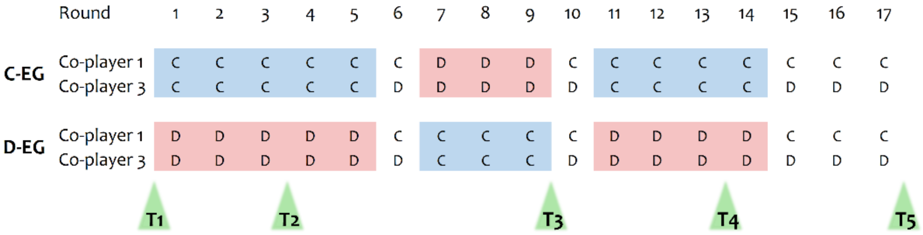
The experiment is presented, and temporal differences of personal and social norm change are analyzed in Batzke & Ernst (submitted). Like the agent-based model, participants play a repeated 3-person Prisoner's Dilemma Game, themselves being one of the three players. Their artificial co-players were predefined behavioral sequences. The behavior of the co-players differed in the two experimental groups (see Figure 1). The cooperative experimental group (C-EG) is characterized by predominantly cooperative social setting and the defective group (D-EG) by a defective setting. The experimental variation was expected to influence the norm internalization process. Moreover, the social setting was varied repeatedly within each group (see Figure 1). This was expected to show in changing social norms. In total, the game consisted of 17 rounds.

N = 365 participants were sampled via a survey institute and invited to play the online "Concert Game". Participants were asked to imagine themselves being a pianist, preparing for the first grand concert. To practice for the concert, they have rented a practice room with an identical piano for 3 hours daily. However, the room is in a triangle with two other practice rooms, and their thin walls make it difficult to practice loudly without disturbing each other. While the pianos have headphone options to avoid disturbing others, it limits the learning achievements. Hence, participants must choose every day whether to practice loudly or with headphones, knowing that they will share the space with the same two people in the coming days.

Before the game, participants social value orientation (hereafter called: willingness to cooperate) was assessed via the slider measure (Murphy et al., 2011). The game was

[1] In the model, social injunctive norms are represented as a constant. Therefore, only agents' and participants' social descriptive norms are compared in Chapter 4. For reasons of simplicity, they are referred to as social norms.

explained, the playoff matrix introduced, an example round played, and participants' understanding of the instructions tested. Before, throughout and after the game, at in total 5 measurement time points (see green triangles in Figure 1) participants were asked to rate their personal norms (e.g., "I am deeply convinced that I should play the piano via headphones.") and social norms (e.g., "The others mostly play the piano via headphones.") from 1 "not agree at all" to 101 "absolutely agree" on each two items.



	Round	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
C-EG	Co-player 1	C	C	C	C	C	C	D	D	D	C	C	C	C	C	C	C	C
	Co-player 3	C	C	C	C	C	D	D	D	D	D	C	C	C	C	D	D	D
D-EG	Co-player 1	D	D	D	D	D	C	C	C	C	C	D	D	D	D	C	C	C
	Co-player 3	D	D	D	D	D	D	C	C	C	D	D	D	D	D	D	D	D
		T1		T2						T3				T4				T5

Fig. 1. Participants played the 3-person Prisoner's Dilemma Game with two artificial co-players (1 and 3), themselves being player 2. The game differed between the cooperative (C-EG) and defective experimental group (D-EG). Each experimental group consisted of three phases, characterized by either a cooperation (blue color) or defection (red color) of the co-players, and a final phase, characterized by a mixed setting in which one co-player cooperated and the other defected (white color). In between phases, single rounds of a mixed setting were added to make the game more realistic. Social and personal norms were assessed before the game (T1) and roughly after each phase (after rounds 3, 9, 13 and 17) at T2 – T5.

4 Comparison of Simulation and Experimental Data

To compare simulation and experimental data regarding participants' and agents' behavior, social norms, and personal norms, the agent-based model was modified so that one agent – equally to study participants – can play the social dilemma game for 17 rounds with two predefined behavioral sequences. The experimental design was similarly applied to the model, with agents playing in the same two conditions: the cooperative and defective experimental group. Hence, agents and participants were put in the exact same situations.

To investigate interindividual differences, results were looked at with respect to participants' and agents' willingness to cooperate. DINO agents were categorized into three groups: cooperators, conditional cooperators, and defectors. Categorization was based on certain ranges of their personal value factors according to the agent type descriptions in Batzke and Ernst (2023). Within these ranges, values were randomly drawn for 100 agents per category and condition. Hence, in total 600 model runs (2 conditions x 3 categories x 100 agent draws) were conducted.

Study participants were grouped along their social value orientation (see Murphy et al., 2011) into altruists (*n* = 148), prosocials (*n* = 142), individualists (*n* = 63), and competitives (*n* = 13). Due to the small number of competitives, the category was merged with individualists. The resulting three categories are hereafter, like the agent categories, referred to as cooperators, conditional cooperators, and defectors.

4.1 Behavior

Figure 1 shows study participants' (Figure 2A) and DINO agents' (Figure 2B) cooperative behavior across the 17 rounds of the social dilemma game, depending on the experimental group (cooperative vs. defective) and their willingness to cooperate (cooperators vs. conditional cooperators vs. defectors).

Agents' behavior generally shows to be less reactive than participants' behavior. However, particularly participants and agents categorized as conditional cooperators show similar behavioral developments across time. They are responsive to the social norm change showing around round 10 as well as to the second change around round 15. Study participants categorized as cooperators show a similar behavioral pattern, while defectors are less but still somewhat influenced by the repeated social norm changes. The respective agent types do not show that pattern. Agent cooperators are generally too cooperative.

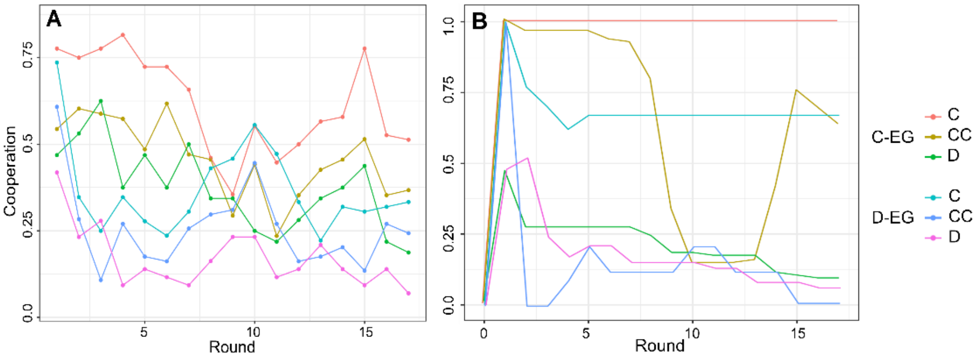


Fig. 2. Study participants' (Figure A) and DINO agents' (Figure B) cooperative behavior across 17 rounds of the social dilemma game, depending on the experimental group (C-EG = cooperative experimental group vs. D-EG = defective experimental group) and their willingness to cooperate (C = cooperators vs. CC = conditional cooperators vs. D = defectors). Cooperation ranges between 0 and 1.

4.2 Social Norm Change

Figure 3 depicts participants' (Figure 3A) and agents' (Figure 3B) cooperativeness of the social norm across time (i.e., 17 rounds), depending on the experimental group (cooperative vs. defective) and their willingness to cooperate (cooperators vs. conditional cooperators vs. defectors).

The patterns of participants and agents' development in social norms matches across experimental groups and willingness to cooperate categories. Yet, agents' social norm adaptation is faster than participants', plateauing after few rounds of the game. Moreover, in participants' the amplitude of adapting social norms to the social setting decays across time, which agents do not mirror.

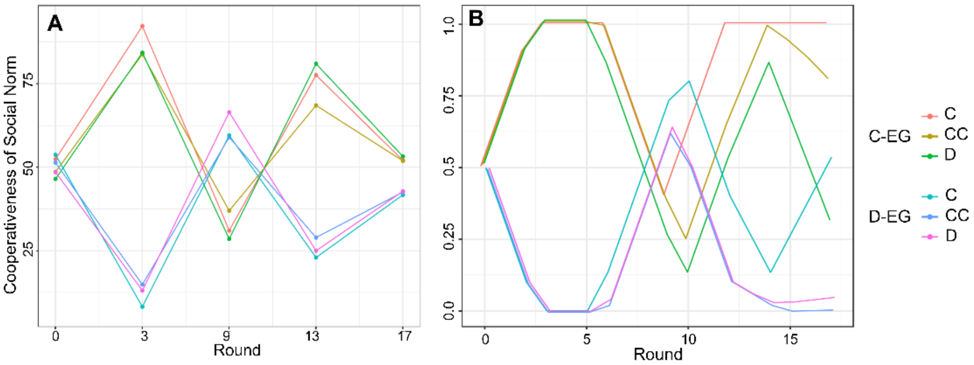


Fig. 3. Study participants' (Figure A) and DINO agents' (Figure B) change in the cooperativeness of the social norm across 17 rounds of the social dilemma game, depending on the experimental group (C-EG = cooperative experimental group vs. D-EG = defective experimental group) and their willingness to cooperate (C = cooperators vs. CC = conditional cooperators vs. D = defectors). In the study (Figure A), the cooperativeness of social norms ranges between 0 and 100, in the model (Figure B) between 0 and 1.

4.3 Personal Norm Change – Norm Internalization

Figure 4 shows participants' (Figure 4A) and agents' (Figure 4B) cooperativeness of the personal norm across time (i.e., the norm internalization process), depending on the experimental group (cooperative vs. defective) and their willingness to cooperate (cooperators vs. conditional cooperators vs. defectors).

When comparing agents' and participants' norm internalization, especially one point strikes the eye: Agents' norm internalization is generally more towards cooperativeness. In participants' personal norm change, there is no learning of a cooperative norm in any group or category, but rather of a defective norm. Nevertheless, the internalization processes of agents and participants that are categorized as conditional cooperators show similarities in both experimental groups.

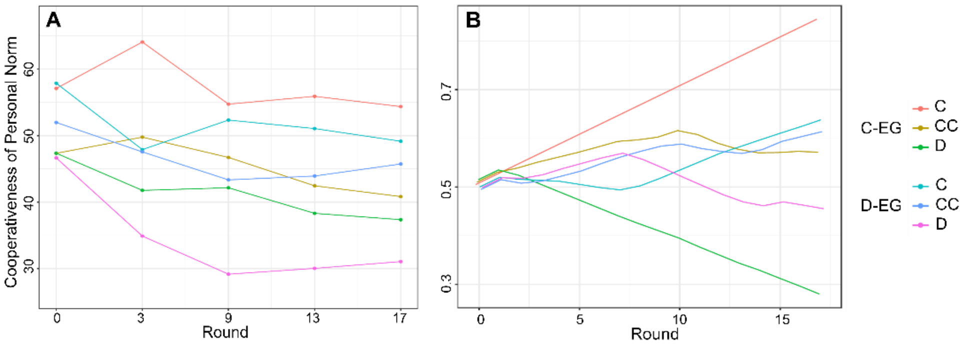


Fig. 4. Study participants' (Figure A) and DINO agents' (Figure B) change in the cooperativeness of the personal norm across 17 rounds of the social dilemma game, depending on the experimental group (C-EG = cooperative experimental group vs. D-EG = defective experimental group) and their willingness to cooperate (C = cooperators vs. CC = conditional cooperators vs. D = defectors). In the study (Figure A), the cooperativeness of personal norms ranges between 0 and 100, in the model (Figure B) between 0 and 1.

The DINO norm internalization process was adjusted to account for that point by introducing asymmetry in the ease of internalizing cooperativeness versus defectivity. The threshold to internalize cooperativeness was raised and thus internalizing cooperativeness made more improbable. Results are shown in Figure 5.

That adjustment significantly improved the overall similarity between participants' (Figure 4A) and agents' norm internalization patterns. Particularly the patterns of cooperators have improved by introducing asymmetry. Regarding defectors, there is still a substantial difference between model and experimental data. In the model, defector agents in the cooperative group (green line) internalize a defective personal norm more quickly than those in the defective group (pink line). In participants (see the same lines in Figure 4A), it is the other way round.

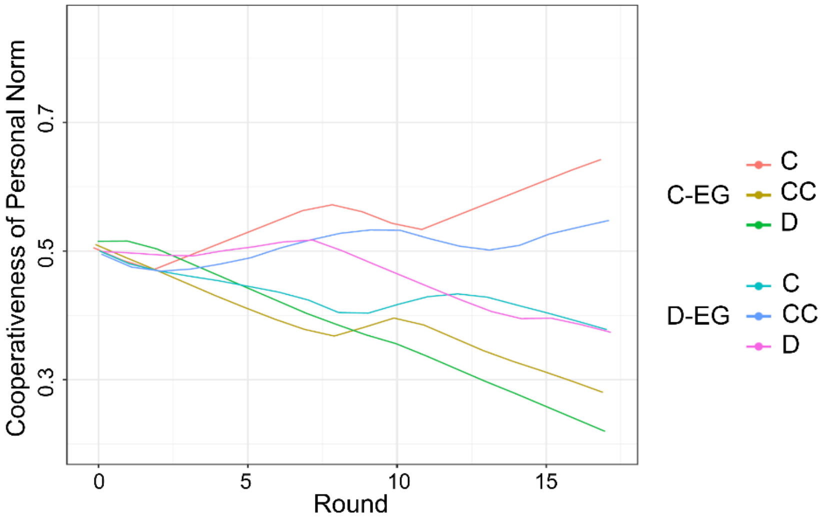


Fig. 5. DINO agents' change in the cooperativeness of the personal norm after implementing asymmetry in the ease to internalize cooperativeness versus defectivity (further explanations in the text). Results are shown across 17 rounds of the social dilemma game, depending on the experimental group (C-EG = cooperative experimental group vs. D-EG = defective experimental group) and their willingness to cooperate (C = cooperators vs. CC = conditional cooperators vs. D = defectors). The cooperativeness of personal norms ranges between 0 and 1.

5 Discussion

The present work represents a first approach at comparing time-series simulation and experimental data on norm internalization. It aimed at validating, testing, and improving an agent-based model on decision-making and norm internalization as well as better understanding social norm change and norm internalization.

Throughout the comparisons of behavior, social norm change, and norm internalization, DINO agents and study participants categorized as conditional cooperators showed considerable similarities, making the DINO model a valuable candidate for further testing and exploration of these processes. The comparison further suggested one mechanism in norm internalization that was so far missing in the DINO norm internalization process: asymmetry in internalizing cooperativeness verses defectivity. Hence, a cooperative norm is more difficult to internalize than a defective norm. The argument of asymmetry relates to Tversky and Kahneman's (1992) *Prospect Theory*. Therein, they describe an asymmetry between losses and gains, stating that negative experiences have a stronger impact than positive. Possibly, this also affects norm internalization.

Implementing that aspect improved the overall similarity of agents' and participants' norm internalization. However, there are several other limitations to the comparison that are not accounted for so far. First, the DINO cooperator and defector agents might be unrealistically extreme. Participants generally exhibited stronger similarities with conditional cooperator agents, which especially showed in the behavioral comparison. Second, the DINO social norm adaptation process is unrealistically fast. Moreover, it does not account for the decay in the amplitude across time found in participants' social norm adaptations. This suggests a decreasing social norm adaptation speed across time. Third, the DINO internalization process in defectors facilitates (rather than impedes) learning a defective personal norm in a cooperative setting. In the model, defector agents may exploit others, which leads to goal fulfillment and thus supporting their actions via internalizing the according norm. While this principle seems to explain some dynamics in the norm internalization dynamics, potentially another factor is missing that accounts for participants categorized as defectors learning a defective norm particularly in the defective condition. For instance, (mis)trust in the others could explain these differences. Mistrust might grow with defection of others as well as repeated behavioral changes.

Comparing internalization patterns of experimental and model data allows investigating key mechanisms that produce observed patterns. Though the present approach provides valuable insights, it needs further research comparing norm internalization times-series data to generate tangible knowledge.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211.

Andrighetto, G., Villatoro, D., & Conte, R. (2010b). Norm internalization in artificial societies. *AI Communications*, *23*(4), 325–339. https://doi.org/10.3233/AIC-2010-0477

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and Applied, 70*(9), 1. https://doi.org/10.1037/h0093718

Atkinson, J. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*, 359–372.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*(4), 1095-1111. https://doi.org/10.2307/1960858

Batzke, M. C. L., & Ernst, A. (2023). Conditions and Effects of Norm Internalization. *Journal of Artificial Societies and Social Simulation, 26*(1), 1–31. https://doi.org/10.18564/jasss.5003

Batzke, M. C. L., & Ernst, A. (*submitted for publication*). Changing fast, changing slow: Investigating temporal differences between social and personal norm change in a social dilemma game. Center for Environmental Systems Research, University of Kassel, Germany.

Bamberg, S., & Möser, G. (2007). Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of psycho-social determinants of pro-environmental behaviour. *Journal of Environmental Psychology, 27*(1), 14-25. https://doi.org/10.1016/j.jenvp.2006.12.002

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social Norms. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/social-norms/

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015-1026. https://doi.org/10.1037/0022-3514.58.6.1015

Dawes, R.M. (1980). Social dilemmas. *Annual Review of Psychology, 31*(1), 169-193. https://doi.org/10.1146/annurev.ps.31.020180.001125

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution, 2*(3), 265–279.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*(3), 629. https://doi.org/10.1037/h0046408

Fishbein, M. & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Boston, MA: Addison-Wesley.

Gintis, H. (2004). The genetic side of gene-culture coevolution: Internalization of norms and prosocial emotions. *Journal of Economic Behavior and Organization*, *53*, 57–67.

Hopper, J. R., & Nielsen, J. M. (1991). Recycling as altruistic behavior: Normative and behavioral strategies to expand participation in a community recycling program. *Environment and Behavior, 23*(2), 195-220. https://doi.org/10.1177/0013916591232004

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making, 6*(8), 771-781. https://doi.org/10.1017/S1930297500004204

Neumann, M. (2010b). Norm internalisation in human and artificial intelligence. *Journal of Artificial Societies and Social Simulation, 13*(1), 12. https://doi.org/10.18564/jasss.1582

Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W.N., Arrow, K. J., Barrett, S., Carpenter, S., Chapin III, F. S., Crépin, A.-S., Daily, G., Ehrlich, P., Folke, C., Jager, W., Kautsky, N., Levin, S. A., Madsen, O. J., ... De Zeeuw, A. (2016). Social norms as solutions. *Science, 354*(6308), 42-43. https://doi.org/10.1126/science.aaf8317

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives, 14*(3), 137-158. http://www.jstor.org/stable/2646923

Otto, S. & Kaiser, F. (2014). Ecological behavior across the lifespan: Why environmentalism increases as people grow older. *Journal of Environmental Psychology*, *40*, 331–338

Sherif, M. (1936). *The psychology of social norms.* Harper.

Thøgersen, J. (1999). The ethical consumer: Moral norms and packaging choice. *Journal of Consumer Policy*, *22*(4), 439-460. https://doi.org/10.1023/A:1006225711603

Thøgersen, J. (2006). Norms for environmentally responsible behaviour: An extended taxonomy. *Journal of Environmental Psychology*, *26*(4), 247-261.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297–323. https://doi.org/10.1007/BF00122574

Villatoro, D., Andrighetto, G., Conte, R., & Sabater-Mir, J. (2015). Self-policing through norm internalization: A cognitive solution to the tragedy of the digital commons in social networks. *Journal of Artificial Societies and Social Simulation, 18*(2), 2. https://doi.org/10.18564/jasss.2759