New England Journal of Public Policy

Volume 36
Issue 1 *The Changing Character of War and Peacemaking*

Article 6

6-16-2024

Employing Multi-Agent AI to Model Conflict and Cooperation in Northern Ireland

Katherine O'Lone
The University of Manchester

Michael Gantley
CulturePulse; Linacre College, University of Oxford

Justin E. Lane
CulturePulse; Institute of Ethnology and Social Anthropology, Slovak Academy of Sciences

F. LeRon Shults
CulturePulse; Institute for Global Development and Social Planning, University of Agder; NORCE Center for Modeling Social Systems

Employing Multi-Agent AI to Model Conflict and Cooperation in Northern Ireland

Katherine O'Lone
The University of Manchester

Michael Gantley
CulturePulse; Linacre College, University of Oxford

Justin E. Lane
CulturePulse; Institute of Ethnology and Social Anthropology, Slovak Academy of Sciences

F. LeRon Shults
CulturePulse; Institute for Global Development and Social Planning, University of Agder; NORCE Center for Modeling Social Systems

Abstract

In this article, we outline the development of a multi-agent artificial intelligence (MAAI) model for post-conflict Northern Ireland. We discuss the insights it provides into the primary drivers of conflict and cooperation in the post-Agreement era. Analyses reveal that leading drivers of cooperation in the model are fairness and sadness, while the main drivers of conflict are related to anxiety and perceived moral authority. We examine these findings in the context of previous computational modeling efforts in Northern Ireland, the social psychological literature on intergroup conflict, and the current geopolitical landscape. We conclude by advocating for the application of this technology as a tool to inform policymaking and address the ethical considerations raised by its use in peacebuilding and reconciliation efforts.

Katherine O'Lone is a behavioral science fellow in the School of Social Sciences at The University of Manchester.
Michael Gantley is the Chief Project Officer at CulturePulse, a US-Slovak company based in Bratislava.
Justin E. Lane is the Chief Executive Officer of CulturePulse, and a research fellow at the Institute of Ethnology and Social Anthropology, Slovak Academy of Sciences, Bratislava.
F. LeRon Shults is the Chief Research Officer at CulturePulse, a professor at the University of Agder, and a research professor at the NORCE Center for Modeling Social Systems.

This research was supported by a grant from a funder who wishes to remain anonymous. The project was led by the first author while a research fellow at the Woolf Institute.

Background

The Good Friday or Belfast Agreement (hereafter the GFA) was signed on April 10, 1998. This historic document marked, for the most part, an end to thirty years of ethno-political violence in Northern Ireland known colloquially as 'the Troubles.' More than 3,000 people were killed and tens of thousands injured. At the time of writing, the international community celebrates the Agreement's twenty-fifth anniversary; rightly fêted as a paradigmatic model of overcoming seemingly intractable conflict, an astonishing feat when one considers that post-conflict peace, on average, lasts only seven years and roughly sixty percent of conflicts reoccur.[1]

However, an agreement is never the end of peacemaking and the resulting peace in Northern Ireland has, by no means, been without difficulties. Political and sectarian threats remain, and more recent issues such as Brexit and ongoing uncertainty surrounding the constitutional future of Northern Ireland jeopardize the relative stability.[2]

In recent years multi-agent artificial intelligence (MAAI) modeling and simulation tools have increasingly been used to address challenges related to intergroup conflict and to explore the conditions for social cohesion and peace within and between human groups.[3] Why have such tools become so popular? Unlike traditional methodologies, MAAI modeling can show the causal links between micro-level behaviors, meso-level interactions, and macro-level emergent social patterns, because what sets it apart from traditional game theoretic agent-based modeling or machine learning techniques is the utilization of psychologically realistic cognitive architectures embedded within realistic social networks. Moreover, it provides stakeholders and change agents with a sort of virtual laboratory, an "artificial society," in which they can test their hypotheses and run scenario simulations before trying them out in the real world. This is particularly important in regions or countries such as Northern Ireland, where peace and social cohesion are fragile and policy makers must move carefully and wisely.

In this article, we outline how the use of MAAI technology can provide policymakers with a powerful set of digital tools to model and predict both conflict and cooperation. We begin by providing an overview of the emergence of MAAI in policy and highlight its explanatory potential. We discuss previous modeling work on intergroup conflict and reconciliation in Northern Ireland, which provided the foundation for our research. We then outline our methodological approach, including the use of sentiment analysis and the creation of a 'digital twin' to simulate the conditions for social stability (or not) in Northern Ireland with a focus on the implications of removing the 'peace walls.' Finally, we advocate for the use of MAAI in peacebuilding and conclude by addressing some of the ethical concerns that arise from this approach in policymaking.

Computational Simulation and Artificial Societies

To model the complexity of human moral and social behavior within artificial societies, social science has used two approaches.4 The first, older approach is grounded in evolutionary game theory and has often been used to model phenomena such as the emergence of cooperation in networks or the influence of socio-cognitive biases in the development of social norms.5 While no doubt useful, the agents in these models lack psychological realism.6 Arguably, the most policy-relevant issues, such as extremism and the willingness to fight and die for one's group, are shaped by sacred values and specific forms of group alignment rather than rational choice.7

The second, more recent approach is MAAI, which creates artificial societies that are populated with agents who are psychologically realistic, complex, and emotionally motivated; they do not always act with rational self-interest. The agents in these models are programmed with

algorithms designed to mimic evolved human cognition, such as the tendency to detect intentional forces (e.g., agents) and to protect ritual coalitions in the face of perceived threat.[8] The psychological realism of MAAI models holds extraordinary promise for policymaking and can provide decision-makers with the explanatory power and predictive insights needed to tackle some of the most pressing problems of our age such as environmental threats on human social behavior and the resolution of intergroup conflict.[9]

This approach is arguably the most powerful tool that the social and computer sciences have for linking multi-level factors in the analysis of complex adaptive social systems.[10] A MAAI approach can, for example, explain the emergence of macro-level social risks (such as conflict and the growth of radicalization networks) by "growing" them bottom-up from micro-level agent behaviors (e.g., responses to anxiety, similarities of belief, and perceptions of the violation of key socially shared values) and meso-level interactions (such as interactions taking place between agents in a social network, online or offline).[11] The following analogy is often useful: more than a "snapshot" of correlated variables at a particular time and space, MAAI provides a "video" of longitudinal causal dynamics that can be rewound and played again under different conditions.[12]

MAAI and Policy

Where MAAI is of most use to the policy realm is its ability to construct artificial societies or 'digital twins' of communities and networks where agents interact with one another in ways that potentially affect the values of other agents and their environment, thus giving rise to population-level phenomena. From here, one can adjust the parameters of the digital twin to observe the subsequent individual and population-level effects of implementing new policies within this simulated environment. These digital twins serve as virtual laboratories in which policymakers can explore and discover the conditions under which—and the mechanisms by which—individual and social variables change in the artificial society.[13] However, the true power of this approach lies in its ability to gauge the likely outcomes of interventions in a simulated, low- to zero-risk environment, before their implementation.

The ability of these models to formulate the complexity of human social systems can offer a valuable way to gain insights into some of the most complex policy issues of our time. For example, MAAI has been applied to modeling emotional contagion surrounding COVID-19 and shed light on some of the mechanisms by which misinformation, stigma, and fear spread throughout Scandinavia early in the pandemic;[14] the rise of nationalism and perceptions of threat during the pandemic;[15] the growing secularization of societies;[16] and the integration of Syrian refugees into Dutch society.[17] Recently, policymakers have turned to MAAI and simulation to assist them in addressing the perennially sensitive challenges of predicting and mitigating intergroup conflict and exploring the conditions needed for peace and reconciliation.

Previous MAAI Model of Religious Violence in Northern Ireland

The research outlined in this article was based on a successfully calibrated and validated model developed by two of the authors (F. LeRon Shults and Justin E. Lane) with other colleagues in 2018.[18] This model of mutually escalating religious violence (MERV) was developed with key social psychological theories (e.g., terror management theory, social identity theory, and identity fusion theory) built into the causal architecture to determine the mechanisms underpinning religious, intergroup conflict. This model was validated using data related to the Troubles in Northern Ireland and the 2002 Gujarat riots in India. Both conflicts were extreme, reaching levels

of severe physical violence yet taking place on markedly different time scales. Despite this difference, both contexts are examples of what the authors refer to as 'mutually escalating xenophobic anxiety' that led to significant violence and the breakdown of social cohesion.

The model was validated at the individual or 'micro'-level in relation to experimental data from social psychology and at the macro-level—that of the emergent phenomena of concern (here the mutually escalating conflict between two groups)—in relation to data from the conflict in Northern Ireland. Using this data and the insights generated from the key theories, the model highlighted the conditions under which the behavior and the interactions of individual agents can lead to mutually escalating xenophobic anxiety.

The most common conditions for this to occur were a) those in which the size difference between the two groups was not too large and b) that the agents experienced social and contagion hazards at levels of increased intensity that passed their tolerance thresholds. It is under these conditions that agents encounter outgroup members more regularly and perceive them as threats, thus generating mutually escalating xenophobic anxiety.

Methodology

Our more recent investigation into the dynamics of conflict and cooperation in Northern Ireland was conducted in two stages. Initially, we conducted a sentiment analysis that analyzed more than thirteen million news articles (from 1979–2022) related to the conflicts in Northern Ireland, extracting the core psychological and moral dimensions that formed the basis of community tension and conflict. Following this, we employed MAAI techniques to identify the fundamental moral concerns driving both conflict and cooperation. This was achieved by constructing a computational model—a digital twin—that simulates the complex interplay of factors influencing stability, cohesion, and conflict in Northern Ireland.

Sentiment Analysis

News metadata was gathered from GDelt (the largest database of human society ever created).[19] Using a text-analytic system designed by CulturePulse, this data set was coded for more than ninety aspects of culture and psychology such as moral foundations.[20] This coding was done using an artificial intelligence (AI)-powered natural language processing algorithm that matches key linguistic markers drawn from empirical studies of morality and psychology with similar markers found within the text of a news article.

According to moral foundations theory there are (at least) six psychological systems that provide the foundations upon which human cultures then build and develop narratives, moral frameworks, and virtues (care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, purity/degradation, and liberty/oppression). The existence of these foundations is supported by a wealth of cross-cultural empirical and experimental evidence.[21] All six moral foundations described in moral foundations theory are included in this AI system's analysis. In addition, our analysis added tracking for the concepts "forgiveness" and "revenge." Alongside moral foundations, the system also analyzed elements of social values (e.g., friendship, family ties, and race/ethnicity), evolutionary threats such as contagion, predation, and natural disasters, and text patterns such as readability and information entropy.

These data were collected and aggregated on a monthly basis to develop a comprehensive longitudinal data set. The metadata of articles available in GDelt includes specific codes assigned to news reports to categorize the type of events they cover. Our team created a further classification

by aggregating these coded events to look not just at the classes of conflict but also at multiple dynamics important to peace and policymaking. Ultimately, the analysis coded all events as either 'conflict,' 'move toward conflict,' 'cooperation,' 'move towards cooperation,' or 'defensive event.' Leveraging these categories and the different moral and socio-psychological dimensions, the AI system is equipped to analyze and identify the underlying factors that influence these events, determining whether they are escalating toward conflict or moving toward peace.

Digital Twin Construction

Digital twin construction relies on the integration of multi-agent AI systems, which simulate the interactions between different agents within a complex system to predict outcomes based on varying inputs and conditions. The conditions can be defined through a mixture of variables and rule interactions as well as by defining the initial conditions of a simulation, which should be matched to the real world as closely as possible. The multi-level approach requires interactions between psychological and social environments, with the social environment being measured either through big-data analysis of media or social media data streams or employing representative survey data. This approach is pivotal in modeling social systems and understanding their stability or susceptibility to change.

The development of digital twins involves a comprehensive analysis of the system being modeled, requiring data sets that capture the nuances of agent behaviors and interactions, often drawn from the psychological, sociological, and historical literature in collaboration with subject matter experts and stakeholders on the ground. Our methodology employs real-time analytics to process and analyze large volumes of data from diverse sources to track themes related to conflict, social cohesion, and societal stability without the need for extensive retraining of AI models. Furthermore, the AI models themselves are designed to be isomorphic in relation to different measures utilized in survey data and psychological studies, facilitating the ability to draw close connections between analyses of large-scale social and historical data sets, social media data, news media, transcripts, surveys, and lab-based studies. This capability is crucial for adapting to the rapidly changing dynamics of social systems and ensuring the validity, accuracy, and relevance of the digital twin.

Results

Sentiment Analysis

Alongside the longitudinal data set of news metadata, a further classification was created by aggregating the coded news events to look not just at the categories of conflict and the multiple dynamics important to peace and policymaking. From this, the AI system analyzed and identified the underlying factors that predict increases in conflict or cooperation. We present the findings below.

Digital Twin Analysis

We started our digital twin analysis by first running the simulation under approximately 138,000 conditions. We then utilized a machine learning technique (based in binary logistic regression) to classify each condition as having outputs that result in conflict or no conflict, in order to better understand what variables are most affecting a context's propensity to result in conflict. We found

that the system can classify conditions for conflict with 90.28 percent accuracy (please see Appendix A for a confusion matrix, a measure of the classification accuracy).

From this model, we then looked to find what are the most important features that the system needs to predict conflict (a technique simply known as feature selection). From this, we were able to map the features and rank their importance (see Figure 1). What we found was that the most important features are, by and large, the number of values (in the sense of beliefs and ethical values) being exchanged between agents in the simulation (Num_Values) and how similar or different their beliefs are concerning those values (Cultural_Dissonance_Percent). A graph of the values is presented below.

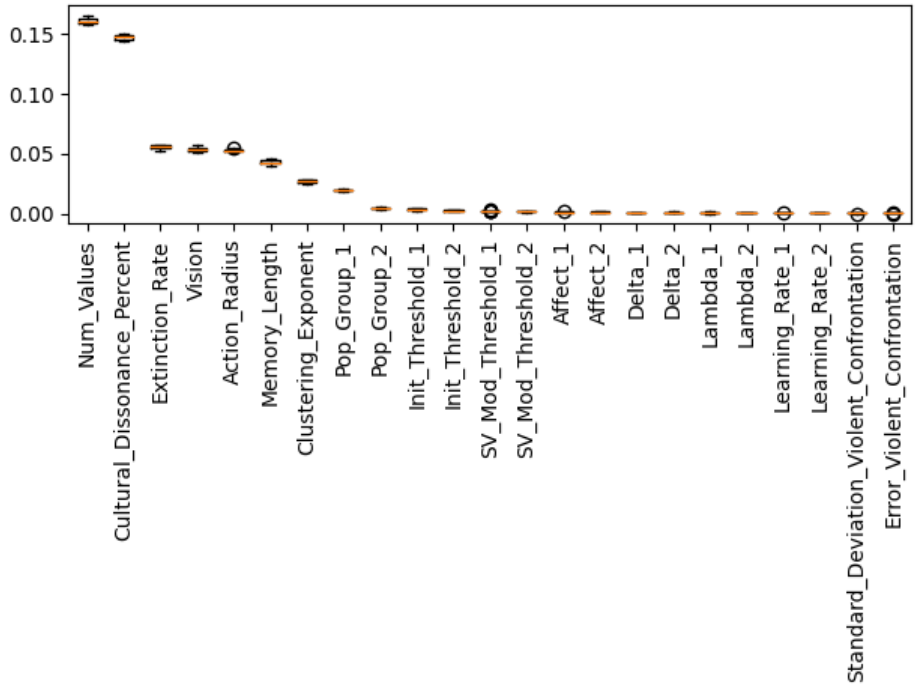


Figure 1. Feature selection: Important features needed by the system to predict conflict. See Appendix B for an explanation of each feature.

Other key variables cover how quickly individuals stop associating negative experiences with the beliefs of others in their environment (Extinction_Rate), how easily they can find others in their environment (Vision), as well as how easily they can interact with those individuals, not just observe them (Action_Radius). This is followed by the Memory_Length, or how far back the memory of each agent in the digital twin goes. In order to simulate how in light of intense emotional experience, there are some things we never forget, the agents in this model also have a form of episodic memory. This simulates the effect of never forgetting—or forgiving—experiences that they might have during a simulation, as in real-world post-conflict contexts.

Discussion of Sentiment Analysis Findings

A quarter of a century on from the Good Friday Agreement what have we found about the mechanisms that influence cooperative and conflict tendencies in Northern Ireland? The results of

the sentiment analysis identified sixty features that underlie episodes of cooperation in Northern Ireland and revealed that the biggest driver of conflict was anxiety. The second biggest factor driving conflict was concerns related to the moral foundation of authority. In contrast, the sentiment analysis revealed that the biggest driver of cooperation in Northern Ireland was concerns surrounding the moral foundation of fairness followed by sadness as the secondary driver. We will now discuss each of these psychological and moral drivers in turn.

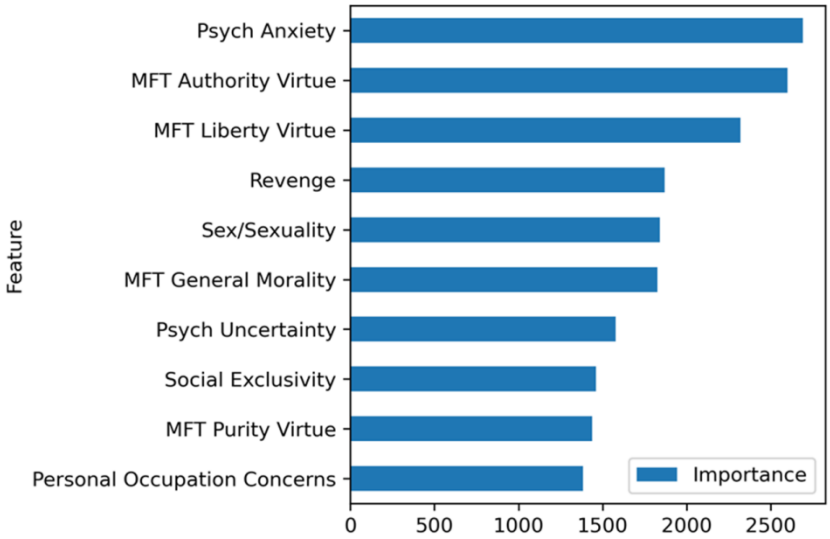


Figure 2. Lead Drivers of Conflict in Northern Ireland. The numbers on the x axis signify importance, with larger numbers signifying greater importance. The features represent anxiety, concerns about moral authority, concerns about violations of liberty, revenge, concerns about sex/sexuality, general moral concerns, psychological uncertainty, concerns about group uniqueness, concerns about violations of moral purity, and concerns about personal/professional resources. According to moral foundations theory (MFT), intuitions about what is moral (or not) rest on at least six psychological foundations (care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, purity/degradation, and liberty/oppression).

Anxiety

The most important feature driving episodes of conflict was anxiety (see Figure 2). This finding corroborates empirical work that demonstrates that anxiety leads to detrimental intergroup relations.[22] Moreover, previous simulation work in Northern Ireland found that episodes of intergroup violence were mostly driven by an increase in the average level of anxiety in the simulated agents over time. One of the most common conditions under which longer periods of mutually escalating anxiety occur are those in which the difference in the size of the hostile groups is not too large. This is particularly relevant to Northern Ireland. When created in 1921, it had a Protestant majority (of approximately two to one) meaning that the Catholic population was the minority group. Over a century, that changed dramatically. Census data from 2021 revealed that for the first time in its history, Catholics outnumber Protestants.[23]

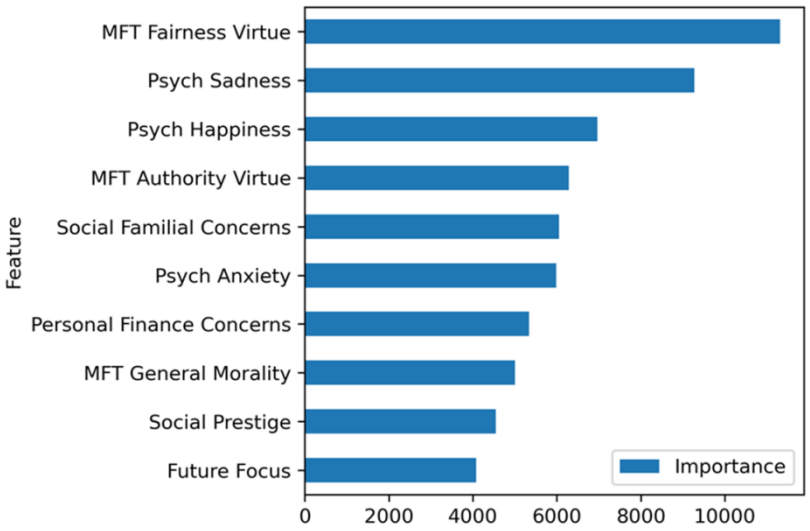


Figure 3. Lead Drivers of Cooperation in Northern Ireland. The numbers on the x axis signify importance, with larger numbers signifying greater importance. The features represent concerns about violations of fairness, sadness, happiness, concerns about moral authority, concerns about family, anxiety, concerns about finance, general moral concerns, concerns about social standing, and concerns about the future.

This is important for several reasons. First, the small size of the difference between the groups (45.7 vs. 43.48 percent) means it satisfies the conditions revealed by Shults and colleagues to be optimal for the escalation of mutual anxiety.[24] Second, the shift from majority to minority status for Protestants (and vice versa for Catholics) will no doubt have a psychological impact. Previous research on projected majority-to-minority demographic shifts has seen increased feelings of threat in members of the once-majority group and a deterioration of positive intergroup attitudes.[25] This demographic shift in Northern Ireland has happened at a time when other political and global events have impacted a strong Protestant or Unionist identity, one that is already alert to threats.[26]

For example, in the aftermath of Brexit, the possibility of a customs border in the Irish Sea has caused enormous controversy; it would psychologically and physically place a hard border between Northern Ireland and the UK, anathema to many Unionists. Moreover, the death of Queen Elizabeth II in 2022 was a seismic change for Loyalists in particular, whose professed allegiance is not to the British government but to the monarchy.[27] With the queen's passing it remains to be seen whether the same emotional attachment to the Crown will endure during the reign of King Charles III.

In addition, close studies of recruitment into Northern Irish paramilitary groups such as the Ulster Volunteer Force have recorded that motivations for enlistment often are rooted in anxiety, although historically it had been an anxiety premised on perceived Irish Republican Army (IRA) threats. So while increased anxieties may lead to resurgences of violence, we cannot expect the same patterns of paramilitarism to repeat in the modern context. The social environment of Northern Ireland and the groups in consideration have changed in how they present themselves publicly, if they present themselves publicly at all.

Authority

The sentiment analysis revealed that discussions in the news around violations of the moral foundation of authority affected the number of episodes of conflict (see Figure 2). This moral foundation evolved in response to our long history of hierarchical and structured societies and in today's world relates to respect for traditions and deference to traditional institutions and authority figures. That these concerns should be driving conflict can also be interpreted in light of the constitutional uncertainty and identity crisis that Northern Ireland is experiencing, particularly from the Unionist perspective. As discussed above, the historical majority identity (e.g., Unionist) is under threat demographically and politically and with that, the legitimacy of traditional authority figures. When a cherished group identity is under threat there is a tendency for members to defend the group and adhere more strongly to shared values and norms.[28] Therefore when authority is perceived to be threatened, this defensive tendency inevitably leads to a breakdown of positive intergroup relations and veers toward conflict. For Unionists, this identity threat is considerable because of the possibility of a border poll in the next decade. If a majority vote to leave the UK, this would lead to a much more pronounced minority status for Unionists; they would be part of a geographically larger united Ireland and exist within a constitutional framework in which both authority and population are not aligned with their social identity.

Fairness

In the twenty-five years since the signing of the Good Friday Agreement, Northern Ireland has, for the most part, remained peaceful. Anecdotal evidence from many of our interlocutors in Northern Ireland, from all sides, emphasizes that although the violence has ended, the signing of the GFA did not address issues surrounding justice and legacy. Now, these dominate the public discourse.[29]

In the social psychological literature, perceptions of injustice or unfairness are often a catalyst to conflict.[30] Our findings suggest that in Northern Ireland, there seems to be a desire for fairness—which moral foundations theory equates with concerns about justice—that is driving and strengthening intergroup relations and moving toward episodes of cooperation (see Figure 3). The importance of careful policy surrounding this feature, given its relevance for cooperation, cannot be understated. State-sanctioned attempts to address the "legacy" of the Troubles have met with a largely hostile reception; the 'Legacy Bill' currently sits in the committee stage in the House of Lords. This bill has been met with hostility from opponents because of a perceived removal of access to justice for victims and relatives of those affected by the Troubles in Northern Ireland. However well-intentioned the bill may have been, it is arguably so unpopular because rather than addressing one of the key drivers of cooperation—issues surrounding fairness—it has done the opposite; people have perceived it as violating fairness concerns.[31]

Sadness

The second largest feature driving cooperation in Northern Ireland is the negative emotion of sadness (see Figure 3). Negative group-based emotions (such as anger) have been previously identified as barriers to reconciliation.[32] In the aftermath of intergroup conflict, all groups collectively experience a set of negative emotions, the most predominant being anger; this presents a significant obstacle to the process of intergroup forgiveness, particularly in Northern Ireland.[33]

However, our results suggest that sadness, a negative emotion, appears to be driving cooperation rather than conflict. The legacy of the conflict hangs heavy over Northern Ireland; it

is inevitable that twenty-five years later people look back with sadness and regret at an episode in their history that caused so much suffering. All groups in protracted conflict perceive themselves to have been harmed or suffered wrong at the hands of the other group. This perception of group suffering often comes with the belief that one's group has suffered more than the other, known as 'competitive victimhood' this can lead groups to view their suffering in comparative terms; it is a barrier to reconciliation and cooperation.[34] However, identifying a common or 'inclusive' type of victimhood (i.e., "we are all victims of the conflict") has been shown to facilitate pathways to reconciliation and intergroup cooperation.[35] This 'inclusive victimhood' (i.e., the acknowledgment that everyone has suffered, regardless of ethnic group) and the accompanying sense of sadness could well explain why it is driving a tendency to cooperate in Northern Ireland.

Digital Twin Findings

Number of Values

The result of the feature selection found that the most important features needed by the system to predict conflict were the number of values being exchanged between agents in the simulation and how similar or different their beliefs are concerning those values. This suggests that if there were a large enough set of differing beliefs within a society, there would be ample alternatives to the "sacred values"[36] that tend to trigger conflict and break down intergroup relations. We must be clear that these are not the only variables that influence conflict in the model and certainly not within the real world, however they provide valuable insights for policy makers navigating the complexity of post-conflict societies such as Northern Ireland.

For example, while increasing the number of values in the digital twin produced lower levels of conflict, it would be prudent to avoid increasing the amount of information in an attempt to achieve this, particularly on social media; this can increase anxiety and depression, two core historical triggers elucidated by the initial media analysis in the first stage of the research.[37] As such, one should be cautioned from leveraging social media to flood a discussion with a greater volume of information in the hopes that it would lessen the likelihood of conflict.

'Peace' Walls in Belfast

The first 'peace wall' was put up in Belfast in 1969 and throughout the Troubles, more than thirty miles of walls were erected to prevent outbreaks of violence between Catholic and Protestant communities. The walls leave a permanent physical reminder of Belfast's violent past, yet ironically they were erected as temporary structures.[38] Many scholars, practitioners, and activists argue that their presence maintains and emphasizes sectarian divisions, making salient the perception of outgroup threats between Catholics and Protestant communities.[39] Public attitudes toward the walls have been inconsistent, with no consensus on what to do with them.[40]

In our digital twin simulations, we varied the extent to which segregation factored into the movements of the agents around their environment to simulate the physical barriers created by the peace walls in the real world. Ultimately, the simulations revealed that when it came to whether this segregationist mechanism had an effect on levels of simulated conflict, we found no evidence that it did. This means that in the model, the continuation of the existence of the peace walls had no effect on conflict.

This result may seem at odds with much scholarship and expert opinion on the peace walls. However, a recent report for the Northern Ireland Department of Justice reveals ambivalent public attitudes toward the walls.[41] For example, fifty-eight percent of people living near a peace wall

reported that their function was to make the residents feel safe. The results of our digital twin simulations largely correspond to one of the main findings of the report: "it is generally expected that if the peace walls are not removed, life will continue on as normal and it will have little impact on their community."[42]

Ethical Concerns and Implications

Scholars and stakeholders are increasingly concerned, and rightly so, about the ethical implications of all forms of AI.[43] However, the issues surrounding MAAI are particularly complex. Which voices are included in the articulation of the assumptions grounding such policy-relevant models? Who decides which simulation experiments to run? What are the dangers that bad actors will utilize such models to exacerbate conflict rather than promote peace? These are valid concerns. We hope to have shown that Northern Ireland is among those complex contexts where the risks and opportunities are so high that it makes sense to apply MAAI modeling in the pursuit of policies for promoting peace.

We must be clear. We do not suggest that MAAI is a 'magic bullet' for policymakers working in reconciliation and peacebuilding. Nor do we suggest that MAAI is without limitations, or that there are no valid concerns surrounding the potential misuse of such technology. However, given that this technology holds such promise for reconciliation and peace in regions scarred by conflict, it might reasonably be argued to be unethical to not use it.[44]

Transparency of Assumptions

Let us start by highlighting one advantage that MAAI does have: transparency of assumptions. An MAAI approach requires that the assumptions built into the architecture of a formal model are made explicit by the researchers from the beginning, as are the purposes of the simulation experiments. What this does, as Paloutzian and colleagues explain, is to make the ethical dimensions and ramifications of the research transparent, giving them central importance and therefore making it much "less likely that they will be used for malevolence or manipulation."[45]

MAAI as a Collaborative and Ethical Endeavor

For an MAAI approach to be successful, we propose that it must be a multi-level, interdisciplinary, and collaborative approach; one that involves equal input from all stakeholders from the beginning of the process. For example, the insights of subject matter experts or those working 'on the ground' can provide valuable contextual insights that help to calibrate and constrain the model.[46] While MAAI can offer a credible way to model human complex adaptive systems, the technology relies on the knowledge and expertise of others to 'flesh out' the simulated society, which ultimately allows it to be used as a tool to assist decision-making.

Ensuring the ethical use and accuracy of digital twins is paramount, especially when dealing with sensitive topics like religion, diverse communities, and social cohesion.[47] Data ethics and digital trust are foundational to our approach as they guide the moral considerations inherent in collecting, analyzing, and applying data in digital twin constructions. Our research aligns with the ethical frameworks provided by the UK's Centre for Data Ethics and Innovation (CDEI) and the Data Ethics Framework guidelines provided by the UK government, as well as the EU's AI Ethics Guidelines for Trustworthy AI and UNESCO's Recommendation on the Ethics of Artificial Intelligence.

Conclusion

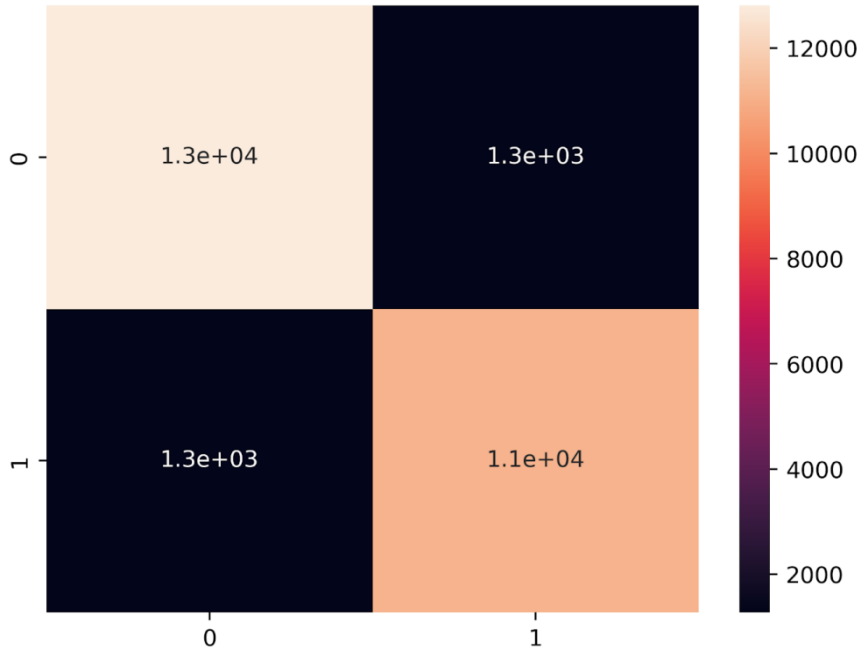
In complex and frequently volatile post-conflict societies, more traditional methods of policymaking and peacebuilding may fall short of addressing the nuanced social and psychological mechanisms that drive communities toward or away from peace. MAAI offers a powerful additional method of testing policy. Our work in Northern Ireland serves as a demonstration of the value of MAAI policy related to conflict and cooperation.

Through the development of a digital twin of Northern Ireland, we have uncovered critical drivers of conflict and cooperation that can assist with the nuanced considerations policymakers must navigate. Our findings suggest that policymakers in Northern Ireland who want to avoid conflict between groups should pay special attention to news events that deal with anxiety and authority, for example. To promote cooperation between groups, policymakers should be mindful of events that evoke a sense of fairness or sadness.

We have also engaged with the ethical dimensions that accompany the deployment of AI in such sensitive contexts. By advocating for transparency and interdisciplinary collaboration, we aim to mitigate potential misuse of this technology. The complexities of modern conflicts and post-conflict societies demand innovative approaches. MAAI, with its ability to simulate and predict the outcomes of various peacebuilding initiatives, stands out as a critical ally in the progress toward enduring peace. Our exploration of its application in Northern Ireland underscores a broader, optimistic narrative for AI's role in peacebuilding.

Appendix

A. Confusion Matrix



Confusion Matrix: a measure of classification accuracy. The numbers on the right signify the number of misclassifications, or errors, from the model. Having high numbers for the upper left hand (0,0), and lower right hand (1,1), and low numbers for the lower left hand (1,0) and upper right hand (0,1) quadrants is good as it signifies that it correctly classified the data more often than not.

B. Explanation of Features in Figure 1.

Num-values. This is the number of values that are being discussed by agents in the environment. It is similar to how diverse the conversation is conceptually.
Cultural_dissonance_percent. How different the beliefs are between two groups.
Exctinction_rate. This is how quickly agents forget about things that happen in the past.
Vision. How close two agents have to be to observe the actions of others.
Action Radius. How close to two agents have to be in order to interact.
Memory_length. How much information agents remember from the past.
Clustering_exponent. How clustered the network is. More clusters mean more subgroups within the network.
Pop_group_1. Number of people in the first simulated group.
Pop_group_2. Number of people in the second simulated group.
Init_threshold_1. How much energy it takes before an agent takes action.

Init_threshold_2. How much energy it takes before an agent takes action.
Sv_mod_threshold_1. How much emotion is required in group 1 before an experience can create a memory that lasts forever.
Sv_mod_threshold_2. How much emotion is required in group 2 before an experience can create a memory that lasts forever.
Affect 1. How emotional the agents are in group 1 initially.
Affect 2. How emotional the agents are in group 2 initially.
Delta 1. This is related to how agents learn and interact based on emotion.
Delta 2. This is related to how agents learn and interact based on emotion.
Lambda_1. This is related to how agents learn and interact based on emotion.
Lambda_2. This is related to how agents learn and interact based on emotion.
Learning_rate_1. This is related to how quickly agents learn and interact based on emotion.
Learning_rate_2. This is related to how quickly agents learn and interact based on emotion.
Standard_deviation_violent_confrontation. This is the standard deviation for how many violent confrontations are observed in any one time period of the simulation.
Error_violent_confrontation. This is a standard error for how many violent confrontations are observed in any one time period.

Notes

[1] Scott Gates, Håvard Mokleiv Nygård, and Esther Trappeniers, *Conflict Recurrence* (Oslo: PRIO, 2016).
[2] Mary C. Murphy and Jonathan Evershed, *A Troubled Constitutional Future: Northern Ireland After Brexit* (Newcastle upon Tyne: Agenda Publishing, 2022).
[3] F. Leron Shults et al., "A Generative Model of the Mutual Escalation of Anxiety Between Religious Groups," *Journal of Artificial Societies* 21, no. 4 (2018): Article 7, https://doi.org/10.18564/jasss.3840; Todd K. BenDor and Jürgen Scheffran, *Agent-Based Modeling of Environmental Conflict and Cooperation* (Boca Raton: **C**RC Press, 2018); Raymond F. Paloutzian, Zeynip Sagir, and F. Leron Shults, "Modelling Reconciliation and Peace Processes: Lessons from Syrian War Refugees and World War II," in *Multi-Level Reconciliation and Peacebuilding*, ed. Kevin P. Clemens and SungYong Lee (New York: Routledge, 2021), 225–42.
[4] Saikou Y. Diallo, F. LeRon Shults, and Wesley J. Wildman, "Minding Morality: Ethical Artificial Societies for Public Policy Modeling," *AI & SOCIETY* 36, no. 1 (2021): 49–57, https://doi.org/10.1007/s00146-020-01028-5.
[5] Rense Corten, *Computational Approaches to Studying the Co-evolution of Networks and Behavior in Social Dilemmas* (Chichester: Wiley, 2014).
[6] Diallo, Shults, and Wildman, "Minding Morality," 51.
[7] Nafees Hamid et al., "Neuroimaging 'Will to Fight' for Sacred Values: An Empirical Case Study with Supporters of an Al Qaeda Associate," *Royal Society Open Science* 12, no. 6 (2019): 181585, https://doi.org/10.1098/rsos.181585.
[8] Justin E. Lane, *Understanding Religion through Artificial Intelligence* (London: Bloomsbury, 2021); Shults et al. "A Generative Model."
[9] F. LeRon Shults and Wesley J. Wildman, "Human Simulation and Sustainability: Ontological, Epistemological, And Ethical reflections," *Sustainability* 12, no, 23 (2020): 10039; Michele Bristow, Liping Fang, and Keith W. Hipel, "Agent-Based Modeling of Competitive and Cooperative Behavior Under Conflict," *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, no. 7 (2014): 834–50, https://doi.org/10.1109/TSMC.2013.2282314.
[10] F. LeRon Shults, "Simulation, Science, and Stakeholders: Challenges and Opportunities for Modelling Solutions to Societal Problems," *Complexity* (2023): 137500, https://doi.org/10.1155/2023/1375004.
11 Justin E. Lane, "Can We Predict Religious Extremism?," *Religion, Brain & Behavior* 7, no. 4 (2016): 299–304, https://doi.org/10.1080/2153599X.2016.1249923.
[12] F. LeRon Shults, "Simulating Supernatural Seeking," *Religion, Brain & Behavior* 9, no. 3 (2019): 262–65, https://doi.org/10.1080/2153599X.2018.1453530.
[13] Shults, "Simulation, Science, and Stakeholders."

[14] Justin E. Lane et al., "Emotional Contagion in Scandinavia during the COVID-19 Public Health Crisis," PsyArXiv Preprint (2024), https://doi.org/10.31234/osf.io/9e5f7.
[15] Josh Bullock et al., "Modeling Nationalism, Religiosity, and Threat Perception: During the COVID-19 Pandemic," *PLoS ONE* 18, no. 4 (2023): e0281002, https://doi.org/10.1371/journal.pone.0281002.
[16] Ivan Puga-Gonzalez et al., "The Rise and Fall of Religion: A Model-Based Exploration of Secularisation, Security and Prosociality," in *Advances in Social Simulation: Proceedings of the 18th Social Simulation Conference, 4–8 September 2023* (Cham: Springer, forthcoming).
[17] Christine Boshuijzen-van Burken et al., "Agent-Based Modelling of Values: The Case of Value Sensitive Design for Refugee Logistics," *Journal of Artificial Societies and Social Simulation* 23, no. 4 (2020): Articl6 6, https://doi.org/10.18564/jasss.4411.
[18] Shults et al., "A Generative Model."
[19] https://www.gdeltproject.org/data.html.
[20] Jonathan Haidt, *The Righteous Mind: Why Good People Are Divided By Politics and Religion* (New York: Pantheon/Random House, 2012); Jonathan Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108, no. 4 (2001): 814–34, https://doi.org/10.1037/0033-295X.108.4.814; Jonathan Haidt and Jesse Graham, "When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize," *Social Justice Research* 20 (2007): 98–116, http://dx.doi.org/10.1007/s11211-007-0034-z.
[21] Mohammad Atari et al., "Morality beyond the WEIRD: How the Nomological Network of Morality Varies Across Cultures," *Journal of Personality and Social Psychology* 125, no. 5 (2023): 1157–88, https://doi.org/10.1037/pspp0000470.
[22] Eran Halperin et al., "Promoting Intergroup Contact by Changing Beliefs: Group Malleability, Intergroup Anxiety, and Contact Motivation," *Emotion* 12, no. 6 (2012): 1192–95, https://doi.org/10.1037/a0028620.
[23] Northern Ireland Statistics and Research Agency (NISRA), *2021 Census* (2021).
[24] Shults et al., "A Generative Model."
[25] Amy R. Krosch et al., "The Threat of a Majority-Minority U.S. Alters White Americans' Perception of Race," *Journal of Experimental Social Psychology* 99 (2022): 104266, https://doi.org/10.1016/j.jesp.2021.104266.
[26] James W. McAuley, "Unionism's Last Stand? Contemporary Unionist Politics and Identity in Northern Ireland," *Global Review of Ethnopolitics* 3, no. 1 (2003): 60–74, https://doi.org/10.1080/14718800308405158.
[27] David Mitchell, "From Queen Elizabeth to King Charles: How Northern Ireland's Unionists Feel about the Monarchy," *The Conversation*, September 21, 2022, https://theconversation.com/from-queen-elizabeth-to-king-charles-how-northern-irelands-unionists-feel-about-the-monarchy-190997.
[28] John Duckitt, "Differential Effects of Right Wing Authoritarianism and Social Dominance Orientation on Outgroup Attitudes and Their Mediation by Threat from and Competitiveness to Outgroups," *Personality and Social Psychology Bulletin* 32, no. 5 (2006): 684–96, https://doi.org/10.1177/0146167205284282; Walter G. Stephan and Cookie White Stephan, "An Integrated Threat Theory of Prejudice," in *Reducing Prejudice and Discrimination*, ed. Stuart. Oskamp (Mahwah, NJ: Lawrence Erlbaum Associates Publishers, 2000), 23–45.
[29] Mark Landler, "In Northern Ireland Town, Painful Memories Lie beneath a Fragile Peace," *New York Times*, April 6, 2022, https://www.nytimes.com/2023/04/06/world/europe/northern-ireland-good-friday-peace.html; Charles M. Sennott, "Northern Ireland's Troubled Peace," *The Atlantic*, May 6, 2023, https://www-stage.theatlantic.com/international/archive/2023/05/northern-ireland-unrest-paramilitary-ira-good-friday-agreement/673969/.
[30] Yochi Cohen-Charash and Jennifer S. Mueller, "Does Perceived Unfairness Exacerbate or Mitigate Interpersonal Counterproductive Work Behaviors related to Envy?" *Journal of Applied Psychology* 92, no. 3 (2007): 666–80, https://doi.org/10.1037/0021-9010.92.3.666; Samuel Fernández-Salinero and Gabriela Topa, "Intergroup Discrimination as a Predictor of Conflict within the Same Organization: The Role of Organizational Identity," *European Journal of Investigation in Health, Psychology and Education* 10, no. 1 (2019): 1–9, https://doi.org/10.3390/ejihpe10010001.
[31] "What Is the Northern Ireland Legacy Bill?" *BBC News*, September 5, 2023, https://www.bbc.co.uk/news/uk-northern-ireland-66648806.
[32] Tania Tam et al., "Postconflict Reconciliation: Intergroup Forgiveness and Implicit Biases in Northern Ireland," *Journal of Social Issues* 64, no. 2 (2008): 303–20, https://doi.org/10.1111/j.1540-4560.2008.00563.x; Michal Reifen Tagar, Christopher M. Federico, and Eran Halperin, "The Positive Effect of Negative Emotions in Protracted Conflict: The Case of Anger," *Journal of Experimental Social Psychology* 47, no. 1 (2011): 157–64, https://doi.org/10.1016/j.jesp.2010.09.011.
[33] Tam et al., "Postconflict Reconciliation."

[34] Masi Noor et al., "When Suffering Begets Suffering: The Psychology of Competitive Victimhood Between Adversarial Groups in Violent Conflicts," *Personality and Social Psychology Review* 16, no. , (2012): 351–74, https://doi.org/10.1177/1088868312440048.
[35] Nurit Shnabel, Samer Halabi, and Masi Noor, "Overcoming Competitive Victimhood and Facilitating Forgiveness through Re-categorization into a Common Victim or Perpetrator Identity," *Journal of Experimental Social Psychology* 49, no. 5 (2013): 867–77, https://doi.org/10.1016/j.jesp.2013.04.007.
[36] Hammad Sheikh, Jeremy Ginges, and Scott Atran, "Sacred Values in the Israeli–Palestinian Conflict: Resistance to Social Influence, Temporal Discounting, and Exit Strategies," *Annals of the New York Academy of Sciences* 1299, no. 1 (2013): 11–24, https://doi.org/10.1111/nyas.12275.
[37] Amandeep Dhir et al., "Online Social Media Fatigue and Psychological Wellbeing—A Study of Compulsive Use, Fear of Missing Out, Fatigue, Anxiety and Depression," *International Journal of Information Management* 40, (2018): 141–52, https://doi.org/10.1016/j.ijinfomgt.2018.01.012.
[38] Jonathon Byrne, "Peace Walls: 'A Temporary Measure,'" *History Ireland* 17, no. 4 (2009): 43, https://www.historyireland.com/peace-walls-a-temporary-measure.
[39] Jack Boulton, "Frontier Wars: Violence and Space in Belfast, Northern Ireland," *Totem: The University of West Ontario Journal of Anthropology* 22, no. 1 (2014): 100–113; Hastings Donnan and Neil Jarman, "Ordinary Everyday Walls: Normalising Exception in Segregated Belfast," in *The Walls between Conflict and Peace*, ed. Alberto Gasparini (Leiden: Brill, 2017), 238–60.
[40] John Dixon et al., "When the Walls Come Tumbling down': The Role of Intergroup Proximity, Threat and Contact in Shaping Attitudes towards the Removal of Northern Ireland's Peace Walls," *British Journal of Social Psychology* 59, no. 4 (2020): 922–44, https://doi.org/10.1111/bjso.12370.
[41] Ipsos MORI Northern Ireland, *Public Attitudes to Peace Walls: 2019 Findings* (2020).
[42] Ibid., 7.
[43] European Parliamentary Research Services (EPRS), *The Ethics of Artificial Intelligence: Issues and Initiatives* (2020).
[44] Nigel Gilbert et al., "Computational Modelling of Public Policy," *Journal of Artificial Societies and Social Simulation* 21, no. 1 (2018): Article 14, https://doi.org/10.18564/jasss.3669.
[45] Paloutzian, Sagir, and Shults, "Modelling Reconciliation and Peace Processes," 237.
[46] Shults and Wildman, "Human Simulation and Sustainability."
[47] F. LeRon Shults and Wesley J. Wildman, "Ethics, Computer Simulation, and the Future of Humanity," in *Human Simulation: Perspectives, Insights, and Applications*, ed. Saikou Y. Diallo et al. (Heidelberg: Springer, 2019), 21–40; F. LeRon Shults and Wesley Wildman, "Artificial Social Ethics: Simulating Culture, Conflict, and Cooperation," in *SpringSim '20: Proceedings of the 2020 Spring Simulation Conference* (2020): Article 38.