AAAI/CCC Symposium on AI for Social Good

Talk Sessions 5: Miscellaneous Applications for AI
Session Chair: Amulya Yadav

Fei Fang: This will be our last technical session on various topics and Amulya Yadav will share the session and Amulya is a PhD candidate at USC, and his research is mainly on influence maximization, especially for homeless youth and with applications for homeless youth and his paper has won the best student paper at AAMAS last year. And yeah, let's welcome Amulya for his opening talk for the session.

Amulya: Thank you for opening, I just want to make sure that all the people who are still gonna be presenting in this session are here, so Yevgeniy, you're gonna be presenting the first paper, okay? And there is Evan Patterson, Oliver has two presentation and mine, okay. So everyone's here, that's great.

All right, so I want to go back, we are nearing the end of the symposium, this is the last talk session and then after this, there'll be an overarching discussion for two hours where we try to synthesize the information that we've gained in the last two sessions, in the last two days and which we will gain in the last session of today. And we try to arrive at conclusions, we try to answer some of the questions that we raised at the beginning of the symposium. And to that end, I will begin that discussion by trying to think about some guiding principles that should be followed or ought to be followed when we are thinking about research AI for social good.

So, over the past two days we sort of looked at this new space of AI for social good. We've seen that it is an interdisciplinary area of research, there is a lot of interest in there and it provides lots of collaboration with several domain experts and several different youth like social work researchers like [Eric 00:02:05], psychologists, criminologists et cetera and more importantly, we've realized that it's not only an application area, it's not the case that you're just gonna apply previously known algorithms to new problems, you are doing things in the field, is actually gonna be able to throw up new fundamental new research challenges that we had not thought of before.

And we talked about in the past doc sessions, we talked about applications in health care, social welfare, urban planning and computation sustainability.

So the ideal case scenario that we have for AI for social good is that there is some real world problem that we go and talk to, that we find out by talking to a domain expert, he tells us about that real world problem, we understand that problem, we build a computer science based model, using that and then we provide an AI based solution and then we give the solution back to the domain expert who then deploys it in the real world. So these problems of AI for social good are characterized by this closure of the loop, so we want to get problems through from the real world and then take them back to the real world to deploy them.

So because these AI solutions that we are building, they are going to be used by human beings, I think it's important to think about some guidelines that should guide the research methodologies that we should follow for AI for social good. And one idea could be that we could use principles from the Belmont report that were highlighted in the Belmont report as a starting point of discussion for principles that should be followed for AI for social good as well.

So what is the Belmont report? How many people know? That's a fair ...It was basically published in 1979 as guidelines or ethical principles for human subject research, it started because of this Tuskegee Syphilis study from 1930s to 1970s where there was a group of researchers who wanted to figure out how is it that the syphilis virus proceeds, progresses in the human body and therefore, what they did was they went to Tuskegee, Alabama and they drew blood from many predominantly black male people without telling them that they have syphilis, so they kept giving them placebo pills without telling them that "You know, you have syphilis, you can be treated for that" And so this is what led to the creation of the IRB, anybody who has worked with human subjects research knows the IRB very well. And this IRB had the Belmont report principles which were followed in that and so I believe this could be a good starting point for our discussion.

There are three primary Belmont principles in the report. They are beneficence, respect to persons and justice. So beneficence says that you want to minimize the risk or possible harms from your research and you want to maximize the benefit of the participant and to the society. Respect for persons says that the research subject who are part of your research should sign an informed consent, they should know what they're getting their foot into when they're being part of your research and justice says that benefits and burdens of research should be fairly distributed among all populations.

Now these are the original principles, what we would like to figure out is what do they really mean in our AI for social good context and I will give you some ideas where these principles might apply and if you have any comments, if you don't agree with those, please feel free to interject.

So, let's talk about the first principle. Beneficence. How does beneficence apply in an AI setting, I would like to give you an example from my own research, so if you attended my talk yesterday, we are doing this prevention, or we are trying to raise awareness about HIV amongst homeless youth in Los Angeles and we will go to these homeless shelters and we try out different AI based algorithms to basically see how much influence do they spread in social networks.

And so we had three different settings that we tried, we had an algorithm called Healer that we talked about yesterday. We tried selecting peer leaders using that algorithm, we tried selecting peer leaders using DOSIM, we tried selecting Degree Centrality based peer leaders and right now, what we're doing is that we're doing an algorithm in order to ensure that it is not the case that random effects into social network are leading to increases and awareness about HIV or getting tested about HIV.

What we're doing is that we have a ... we are recruiting a set of homeless youth and we are not doing any sort of treatment on them, we are not spreading any information amongst them, not doing nothing, any interventions. Now in this setting, does the principle of beneficence apply? Because one could argue, if it were the case that at the beginning of this pilot study when we recruit these homeless youth, if we do find out that the majority of those youth are actually HIV positive, do we abandon this null treatment plan and start giving interventions to those people because they really need or do we go ... because that would be the beneficence thing to do, or do we continue with the null treatment plan [inaudible 00:07:21]

So one could argue that in the long run, finding out which is the correct intervention strategy makes sense because in the long run, you wouldn't want to use a strategy which is suboptimal which could lead to many HIV infections, right? And so in that sense, doing a null treatment makes sense. But if it was the case that you already know that there are many people who are HIV positive, would you still want to do this null treatment?

So I would like to give you an example from cross over study in cancer treatments where each person in the study is basically kept on a placebo for half of the amount of time and on the actual cancer medicine for the other half of the time and this ensures that every person in the study gets cancer medicine. So can similar principles like these be applied to these AI for social good settings?

It was hard for HIV prevention because we were to measure how much information spread, how many people found out about HIV and so if it is the case that we inform a person once about HIV that he cannot be uninformed about HIV later. And so that is something to think about.

The second principle is respect to persons. We should inform people who are going to evaluate the AI system, what is it actually going to do to you. An example that comes to my mind is the following. So you have these smartphones that are highly ... you know, they are powered by AI and one thing that happens in these smartphones is that location information is turned on by default in most of these smartphones. Now how many people in this room, when they bought their smartphones, realized that facebook and google were able to [inaudible 00:08:59] the exact part, they were able to [inaudible 00:09:02] the exact buttons of all the location information that was being shared. How many people? Very few.

And so if it were the case at the beginning of when you don't know your phone, you're going on facebook for the first time, there is a message that pops up that "You know, we're gonna store all this location information for you", I would argue that very few people, substantially fewer people, would opt into this program and so does the principle of informed consent apply in this setting? And I mean, it's not that we're lying to them, you know they can go online and search for this information, it's publicly available but it's not that they're being informed explicitly every time they are acting onto it. And Twitter, Facebook, google all of them are sort of utilizing this information to [inaudible 00:09:46] or recommendations to just improve a user experience.

The third principle and the final principle is justice. It says that benefits and burdens of research should be distributed fairly among all populations of your study. I would like to link it to Sharad's work on pre-trial release and suggest the following question, so what it was doing was that it had a machine learning model and it was releasing people which the machine learning model asked to release.

So in this setting, the benefits of this machine learning model, the benefits of this research are being accrued only by the released people whereas the burdens are only being faced by the people who are kept under arrest. Now how does the principle of justice apply to this setting or an ever broader question is, is justice really the right principle for us to be thinking about in this AI for social good setting. It doesn't have the vermin, so we don't really necessarily have to follow the Belmont principles, there could be another set of principles which apply better to this setting.

So, I guess the question here is what standard of evidence of whether a person should be released or not is appropriate in these settings. When should we release a person, how careful should we be in releasing these people? And is there a principal way of coming up with these evidences?

More generally, with respect to justice, any research that you have in social choice or welfare maximization in which you are trying to fairly allocate resources to people and you're trying to maximize the global welfare as opposed to maximizing the individual utilities of people in that some part of the population is always going to be on the wrong end of the bargain in order to maximize global welfare. And so when you're trying to maximize global utility, some people are going to suffer because of trying to maximize welfare in the entire society. And is this justice for all individuals? So how does the justice principle apply in our setting? So these are some of the questions that I thought of, there could be many more questions.

So to summarize, we have the space, we have problems characterized with this closure of the loop and since the solutions that we're gonna develop are gonna be used by human beings, it is important to think of guidelines to design research methodologies for AI for social good and the Belmont principle provides a first starting point for this discussion.

So how about this, we stop now, we have the talk session, we can all think about these problems, if you have any questions we can take them during the overarching discussion session which will be for two hours so that we can return to these questions later.

Okay, that's it. With that, I'll like to call the first speaker, Yevgeniy.

So Yevgeniy is gonna be presenting optimal thresholds for intrusion detection systems.

Yevgeniy: Okay, thanks everybody for sticking around for the miscellaneous section. This is a talk actually, as you'll see, is somewhat applicable to the urban planning session as well, it's really contextualized in the context of cyber physical systems in urban water networks.

So first of all, intrusion detection systems, generally speaking they are deployed to detect malicious activity, there are different kinds of intrusion detection systems and I'm not going to make that distinction right now, I'm gonna talk about them fairly abstractly.

But the idea is that if you detect some malicious activity, you raise an alarm which then, once the alarm is raised, there is some investigation that goes on and we're gonna mention different examples of this. So for example you can detect suspicious system call sequences, this is in the context of cyber security compromises and monitor system files for modifications and so on.

So the key practical challenge that this talk is going to address is actually somewhat similar to the talk that [Emin 00:13:48] gave earlier, is the imperfections in intrusion detection systems of two kinds. First of all, the false positives, it triggers alarms when alarms should not be triggered and the false negatives which are you don't detect actual attacks. And the problem is of course, both of them have consequences and the consequence is different. It's those consequences we want to capture and trade off between.

So the key way we're gonna address it, is to develop configurations of IDS in terms of optimal detection threshold. Most intrusion detection system essentially give you some form of a score and you can design the threshold, determine which threshold you're gonna operate on and that's going to allow you to trade off between the false positives and false negatives.

Now, in cyber physical systems, one of the major additional challenges is that you don't just get one of these IDS, you get a bunch of them, typically for example in a water network which you see on the right, you have sensors deployed and sensors are not your little sensors but we're talking about computing systems that sense some abnormalities in the water supply. These sensors are deployed in this network and sensors can be compromised for example through cyber attacks and what the intrusion detection systems will do will try to detect these compromisers.

Now, you're gonna deploy a collection of intrusion detection systems which means that now you have this collective information that you can use and the questions is, how you use this information in the aggregate and trade off, again at this higher level, the aggregate level, the false positives and the false negatives.

So the core problem that we're gonna study is finding detection thresholds for multiple, for a set up intrusion detection systems but in the face of strategic attacks which is one of the other major distinction from what [Emin 00:15:34] talked about.

So here is a system model. We assume that there is some cost for investigating false alarms. So any time, anywhere in the space false alarms gets triggered, you have to go and investigate or send somebody to potentially fix something, okay, if it's a false alarm you are wasting some amount for example of money and that's captured by the cost. The other aspect that's important is that the false positives and false negatives induce a trade off. You can't arbitrarily set those things independent from each other. So this is

often conceptualized as this kind of of a curve, there are different ways to draw the curve, [inaudible 00:16:07] is another way. But basically you get to either [inaudible 00:16:10] the false, f of s which is the false negative probability which is going to give you the corresponding false positive rate or vice versa.

So I'm going to assume that we get to choose a collection of false negative probabilities for a collection of sensors and these are [inaudible 00:16:24] by sub s because we get to pick them, potentially in a heterogeneous way for the different IDS systems.

So now what is the attack model. The attacker is allowed to compromise a subset of sensors thinking about in terms of cyber attacks for example. Okay, we call this a, a subset of s, s is the entire collection of sensors that they could compromise. Okay the defender will detect an attack if any of the IDS is [inaudible 00:16:48] along. And otherwise the attack does not get detected.

So if you assume that these alarms are independent which is a reasonable assumption in the setting I described, then the probability of that an attack on a set a is not detected, it's just the product which is what you see here.

Now any time the attacker successfully attacks a subset of sensors that's going to cause a certain amount of damage and you're gonna see in the numerical illustration later what this damage would correspond to, right now, we're just gonna call it some damage function d of a, which is a set value function.

So this induces a game between the defender and the attacker. Utility functions are basically what you would internally expect in this setting, so the attacker's payoff I'll start with cause it's a little simpler but basically it's just the expected damage. So the damage times the probability of that, it doesn't get caught.

Okay the defender's loss captures this as a loss and by [inaudible 00:17:39] it also has the cost term of following up on the false positives and the defender is trying to trade off these two costs and the attacker is basically just trying to maximize expected damage. That's the idea.

So this gameless model does a Steckelberg game where the defender chooses first the configuration settings for all the IDS and the attacker best response by choosing a, the subset of those two attack. Okay, even in the setting where the damage function is submodular, it turns out that even just the attacker's best response to the problem is NP-hard. Fortunately, because in the context of submodular damage functions, it's an unconstrained submodular maximization problem, there turns out to be good approximation algorithms, this is just example of that, it gives a one third approximation in linear time and it's actually quite effective in practice.

The problem of intruder detection threshold setting is considerably harder than that. So we did was, we developed a simulated annealing based algorithm for dealing with this problem which uses the approximate attack I just described as a subroutine.

So in numerical illustration that I'll talk about shortly, we compare it two baseline strategies that people would typically use in this setting. One would be a uniform threshold strategy, basically just computer single threshold for all intrusion detection systems, that's the typical approach. Another one is called locally optimum, you just treat each IDS as an independent IDS, ignoring all the other ones. Those are the two baselines.

So here is the context, you have a water network, you have a collection of sensors that are detecting leakages, the idea is that a sensor can detect a leakage if there is a pipe that's burst within a certain distance from that sensor. So the attacker may temper with the sensors to cause damage here. Damage is basically the undetected leakages and it turns out to be that the damage function is very naturally submodular in this setting because it's based essentially on the sensor coverage of the links.

So we looked at a particular water network with about a 170 pipes and 126 nodes, 18 sensors turned out to be sufficient to cover the entire network here. Everything else I already described. This is what the false positive, false negative trade off curves look like if you use intrusion detection system data sets based on system call sequences. Okay, almost done. And here just the final results, blue line, so lower is better, blue line is our approach and the red and green are the alternatives that I described.

So at this point I think I'm out of time, I'll stop and take questions.

Speaker 4: So it was a little unclear, at one point you were saying attacking the system, at other places attacking the sensor, so I wasn't quite clear what was being attacked.

Yevgeniy: We are modeling cyber attacks on sensor. So sensors being basically ... sensors are just computer systems that are monitoring something and we're just imagining like host based IDS deployed in those systems. I use systems, I guess, [inaudible 00:21:01] meanings, I'm sorry about that.

Speaker 4: So with the Steckelberg game, in order to find the values of the targets, are they values of the systems that those sensors protect or how do you derive the values in the game?

Yevgeniy: So the damage function is really the operative part and that's in terms of the leaks you wouldn't be able to detect for example in the water network. So those sensors are monitoring for leaks, if you compromise a sensor, you can disable it essentially in a way that [inaudible 00:21:29] so cyber attacks are one of the reasons, that's the focus. They are much harder to detect than for example if you actually break a sensor, people will find out pretty quickly. Cyber attack you just [inaudible 00:21:38] catching things, that's the idea. So the damage function is basically physical damage that you incur.

Amulya: Other questions?

Speaker 4: Physical damage that you wouldn't know because you [inaudible 00:21:56]

Yevgeniy: Correct, so the physical damage in the water network you haven't detected, let's say a contaminant or you haven't detected something, a leak or something like that. I mean here in this example it's a leak but we also looked at for example contaminant detection.

Amulya: If there are no other questions let's thank the speaker.

Now we have Evan Patterson from Stanford University and he will be presenting machine representation of data analysis. Towards a platform for collaborative data science.

Evan: Hi, I'm Evan, and so I'm gonna tell you about work that I've done with my collaborators at IBM on the semantic representation of data analysis and potential applications of that for collaborative data science.

So as we've seen from the many wonderful session in this symposium there are many exciting applications of artificial intelligence and data science to social good. We've also seen that there are a number of special challenges that don't arise in traditional applications of AI, I say in like chess playing or go playing.

So it's not taking place in a closed world and as we've seen collaboration between domain experts is essential and not just between them and us but also other people, so policy makers, philanthropists, people on the ground, stakeholders in the problem.

So for example part of what motivated this work is this organization called the Accelerated Cure Project, so what they do is try to stimulate research on multiple sclerosis by providing data on an open access basis to researchers in that field, so that includes analytical data as well as physical biosamples and one of the problems that they face is trying to organize the efforts that are being done on their data repository and being able to disseminate that information and helping the different researchers know what's going on in the field. This is of course not restricted to MS research, it's sort of a general problem that applies when you're working in these complex domains that involve lots of people and data driven questions.

So we had that thought that maybe we could create a cloud platform for collaborative data science that might facilitate these activities. You may ask, well, aren't there already many data science platforms and it's true in essence, but we imagined some features that don't exist in the current offering. So one of the things we'd like to be able to do is have artificial intelligence making recommendations about relevant data analysis or potential collaborators so we talk about ... so where would the collaboration with domain experts come from or do domain experts find collaborative in the data sciences. Well, often it's more or less serendipitous but can you have what's sometimes called design serendipity, a way of helping these interactions to develop.

And then we'd also maybe like to be able to organize analysis in an automated way or evaluate them according to various metrics such as sensitivity analysis or their performance, like held out data and so forth.

So there is a lot there but a common theme that one can see behind this is if one wants to do that you need to have a representation of the content of such a platform, particularly the data analysis that are hosted on it, in a way that is machine interpretable. In a way that our algorithms can work with and so that brings us to the technical contribution of this work which is to create a system for doing this so it will automatically extract a semantic representation of a data analysis. And so I'm gonna tell you a little bit about that.

First, though, let me clarify what I mean by data analysis since this may be not entirely conventional. So I'm thinking of a data analysis as a computer program not so much like a method section in a scientific paper. So here is an example of something I put together, so this is in a notebook form which I'm hoping could become a model for how data analysis is disseminated. So it mixes human text with code, you've probably seen things like this before and there are plots and various things. So this is how I'm conceptualizing data analysis and my system takes this as input.

And what it produces is, I'm sorry, the appearance is not great on this projector but it produces a data flow graph which is showing these steps of this data analysis, at least at a high level. So I'll go through it very quickly, it's just to give you a flavor.

So there is some data being read in, there's some initial group processing, a multiple correspondence analysis model is fit, some transform datas obtained, those are used to fit a [inaudible 00:27:09] clustering model, the clusters come out of that, additional data is read in, they are merged into some larger data frame, finally some plots are produced and if you were to go through this data analysis, you could see that that's what it's doing and it's doing it in python using standard libraries like [Candace 00:27:28] and scikit-learning.

So let me say now in [inaudible 00:27:34] very briefly, how this works and for more details you can look at the paper or ask me afterwards. So this system is based on dynamic program analysis, so at the highest level what it does, it takes in ... it runs the program and traces it and there are several steps to that.

So first, a directed acyclic graph of all the function causes is built up and that is ... at least, all the functions calls it the user code level and that's quite a complicated object and it also doesn't have any semantic content. So at the next step, we have an annotation data base for common statistical software. So you know scikit-learn and libraries like this which identify particular classes and function calls and tags them. And then finally, we have a knowledge base or an anthology of data analysis concepts to which those tagged function calls and objects are aligned.

So this last step is important because it allows for methods which may be instantiated in many different libraries to have a common representation and to be compared. So in this picture showing a small segment of the clustering models that are in our anthology.

Okay, so I think of this as being part of a vision or a dream which certainly I'm not the only one to have of a knowledge ecosystem and the sciences and more broadly, it's not

only open and online but which is somehow ontologically integrated that the knowledge that's embodied in it is represented in a way that can be understood by machines and manipulated by machines and we're a very long way aways from realizing that vision but I think that even partial progress towards it could be very beneficial for our social good and for the scientific progress.

So that's it, thanks for your attention.

Yevgeniy: So this seems really interesting, I wanted to ... I have sort of one question but in two parts. First, to what extent can you do this for procedural but not necessarily program based approaches and this question is motivated by the kinds of protocols you see in computational biology for example and things like [rosera 00:30:22] publications, I don't know if you know about [rosera 00:30:25] , it's a protein modeling tool. When people design new protocols which are combinations of actual code and sequences of steps in which to run the code, and I think that they could really benefit from analysis of this kind if ... but we'd have to account for the fact that it's a mixed procedure.

Speaker 4: So, right now, my system supports python and a planned add support for other common data analysis languages like R and Julia. There is an issue about there is so many tool out there that it is hard to think about supporting everything but I do think that there is a role to be played like you said and even in any moderately complicated data science project, there are tons of steps and files and things floating around and even maintaining the dependencies of those things is challenging and tools like this could facilitate that but in presenting a high level picture of how the different components of the project are related to each other. So I think that's a great point.

Speaker 6: I just had a question about ... so data science is essentially a list of python notebooks nowadays or work flows, right? But any workflow usually is performed because it has a goal and especially if you want to learn what that workflow is about and looking forward if you want to compose different work flows, one would need to represent also inputs, outputs, you know, the whole strategy around it. So you could even analyze whether this workflow is efficient from the point of view of simplicity of achieving a goal or even appropriateness, like [inaudible 00:32:13] that should be applied give in where the ultimate goal is.

Evan: Yes, it's another really good point and so right now, I'm representing knowledge about the data analysis part of this process and we know that this is only a really fairly small part of the larger scientific process or like the social good process and so one of the things that I would really like to do in future work is figure out how to integrate with the domain specific anthologies that may exist, so the biologies for example have done a lot of work towards creating these anthologies. And other fields are in varying places on that but being able to say "No, like okay, I loaded in this data frame and this column actually, this is a gene expression data and it comes from this place and I can track it through this analysis and tell you where it's coming out on the other end", things like that, I can't do yet but I think is one of the most important directions for future research work on this topic.

Speaker 6: [inaudible 00:33:14] the two data sets are used which are not normalized and cannot actually be combined, also common [crosstalk 00:33:20]

Evan: Right, being able to detect those kinds of problems. Yeah.

Amulya: Any other questions? Let's thank the speaker.

Next up we have Erisa Karafili and she'll be presenting argumentation based security for social good.

Erisa: Hi everybody, I'm Erisa Karafili I'm a research associate at the Imperial College, London, I'm working at the Resilient Information System Security group, I'm here for presenting the work Argumentation by security for social good which I have done in collaboration with Tony Karkas, Nico Spanudakis and Emil Lupu. During this work we were supported by two projects. One is a European project [inaudible 00:34:03] cloud and the other is a UK national research project called [Cpart 00:34:10].

I'll start by first giving an introduction about the solution that we're gonna use and the two problems that we will solve with these proposed solution. One is the attribution problem in cyber attack and the other is the data sharing agreements.

So we address two important problems in social context, attribution in cyber attacks and regularitory data sharing. This problem seems that have nothing in common with each other rather than they are both secret problems but actually they can both be seen as decision making problems where the decisions are taken under incomplete information and also conflicting one.

This is why we decided to use argumentation reasoning which is a well known technique for taking decisions under partial conflicting and context dependent knowledge. Because we have this conflicting information it's common to have different kind of conflicts and we are able to capture these conflicts. We are not just able to capture the conflicts, but also to solve them.

So the first problem that we solve is attribution problem with cyber attacks. What is attribution? In a cyber attack is finding out who did the attack. So it's the process of assigning an action to a particular entity, country, actor.

Attribution is important, is very important nowadays where cyber attacks are increasing. Cyber attacks are increasing due to the increase of interconnectivity or to the expand use of IoT devices. During the talks in these two days I have noticed that in social goods, we are using a lot of IoT devices, they are becoming essential nowadays but I also see this kind of trade off that for not risking the security of the data we cut out the device. So the device can not go on internet because we don't know how to insure the security of this device but by doing this, we are losing efficiency and all the different results that we can get by making this device real and IoT device.

Attribution is important because by knowing who did the attack we are able to secure the system, we are able to put into act efficient counter measures, we are able in the future to diminish this attack but also to bring the person that did the attack into justice and this is very important. So attribution is not trivial, there exists no complete theory until now that is able to decide who did the attack, this is also due to the incomplete and conflicting information. Forensics helps in collecting the information but until now, this process is done manually by the analyst.

This is why we decided to propose a solution that is based on non-monotonic reason that automate this process and gives an important help to the analyst itself. The methodology is based on abductive and argumentation reasoning, we construct an attribution reasoner which use logical rules. This rules are [inaudible 00:37:27], how the analyst behaves when he needs to decide who is the person that did the attack, we use the Q model that is a social model taken from war science to categorize and collect the various evidence and we are able not just to say this is the machine that did the attack but we are able to say, this is the group of attackers that performed the attack and why they wanted to perform this attack. If we have the information, we are gonna point it out.

So once we have collect all that, the evidence can be seen as facts or defeasible knowledge. The rules are seen as arguments, as these arguments are conflicting for deciding whose argument that wins, we put hierarchies between that. How we decide who is the argument that is going to win depends on the context and for sure, we also give an explanation of every time we are taking a decision. We have implemented this reasoner by using the GorgiasB tool.

The second problem that we solved is data sharing. Data sharing is playing an important role in society, we are doing our work, we always sharing information, we are sharing information during education, e-health and so on. So think about a patient that is in a remote area and he needs also health care. We can provide this health care by having this remote ... a doctor that is in a new clinic that is ... he can give a remote visit to this patient. The patient can have this IoT devices that can measure the blood pressure, can make analysis, the data are sending to the cloud, the doctor can check the data by the cloud and maybe change the prescription of the patient and so on. While doing this, we still want this patient to be able to ensure the security of his data and that his data are used appropriately.

This is why, before the data is shared, we create an agreement. This agreement is seen as some kind of contract where we had the data security requirements, all the user requests, so how he wants the data to be handled, the business rules of the different entities that are involved, the clinic, the hospital and so on and also the legislation rules. Deciding the rules that can be applied in particular cases it's not easy. One because all these rules are heterogeneous, two because we can have different rules that can be applied on the same case. Not just normal rules but also legislation rules, these are the most hard to be able to first to represent, two to deal with and to decide which rule we can apply for a particular case.

So we propose a decision process model based on Argumentation, we have implemented it in e-health example. Our decision process is able to decide the rules that can be applied and so one, also to decide who and how can access to the data, if you can share the data or send the data to another entity and so on.

These rules are called policies, we are actually the first one to use this technique on policy analysis that is quite known in security. The decision for sure I raised on context and we are able to represent this kind of decisions, we are able even in this case to capture the various conflicts that we can have and also to solve them by using priorities.

We have two concrete applications, one is Coco Cloud from the European project that we have implemented these decisions in various hospitals in Spain and also MEDICA that is an application from Cyprus that they have implemented all the various legislation and rules of Cyprus and also the European Union Community.

So as conclusions, we presented a solution for two different problems that we think these problems are important for social good. One is the cyber attack attribution and the second is regulatory data sharing. The solution is based on argumentation reasoning and we proposed a decision making mechanism under incomplete, conflicting and context dependent knowledge. As future work, we would like to hae quantitative arguments strength, we would like to extend attribution solution to guide the analysts during the evidence collection so not just say, "He did it", but also to say, "Hey we have some information here, maybe go and ask some other questions or go and check this other evidence." To work on human cognitive reasoning for the social evidence and to have a fully automate conflict resolution.

Thank you.

Speaker 8: I was wondering how you could use quantitative argument in your analysis?

Erisa: So until now we see a certain argument is just more stronger, while we would like to put some degree on how strong this argument is. So in this way we are just saying yes or no, he wins, or he lose while we want also to put quantitative, so to put a grade on how much he wins in a way at the end to have a quantitative decision. So not just say, "Okay, he did it" but "He did it with this grade of decisions"

Speaker 8: And then you define conditions over these-

Erisa: Yes, everything depends on the various conditions that you have.

Speaker 9: Interesting talk. So non-monotonic logics are notorious for leading to paradoxes, they don't have a sound and complete-

Erisa: I'm sorry, I cannot hear you.

Speaker 9: I said, non-monotonic logics don't have a sound and complete reasoning system, right? So in order to make this work, you obviously had to make some restrictions, so I was wondering if you could say a few things about that?

Erisa: So not all non-monotonic reasoning have .. are not sound ... So for sure they are sound, not all of them are not incomplete. In this case-

Speaker 9: You need restrictions but-

Erisa: Yes, you need restrictions and yes we-

Speaker 9: [inaudible 00:43:44]

Erisa: I know but in this case you just need to add your restrictions and you can have completeness. So what we are doing until now is that especially with abduction, we are trying to get the best solution for the information we have because we are still under incomplete information. So for the solution, for the information that we have in this exact moment, this is the best solution that we can get. If further information is added then-

Speaker 9: So you have an optimization function that is actually looking at.

Erisa: Yes.

Speaker 9: All right. So that was the restriction essentially.

Erisa: Yes.

Amulya: Let's thank the speaker.

And now we have our final talk of the symposium, we'll hear a talk by Mahendra Prasad, back to the future, a framework for modeling ultra stick intelligence explosions.

Mahendra: Saving the worst for last. In any case, so I'm talking about frameworks for intelligence explosions, so there are several people today who have cut forth models for intelligence explosions, people like Ray Kurzweil, he is a director of engineering at google and Nick Bostrom at the Future of Humanity Institute and while there is several different predictions, some of the predictions include the notion that intellectual capacities will increase indefinitely into the future and that human life spans can be extended indefinitely into the future. And so these are relatively recent, since at least the 1990s predictions that had been made.

But Condorcet did it first. So Nicholas de Condorcet was an 18th century French mathematician, he was championed by his teacher d'Alembert and he made both those predictions in the last chapter of his non-technical philosophical work Sketch of an Historical Picture of the Progress of the Human Mind, that's from 1794, 1795.

What has generally gone unnoticed is that he had a mathematical model for these claims and his Essay on the Application of Analysis to the Probability of Majority Decisions and that's from 1785 and that text doesn't have a complete English translation.

So the simplest version of this Jury Theorem, the political science version of the Jury Theorem is you have a statement s, it's one of two states, true or not true. You have m agents and you have a knowledge condition that you know. Each agent has a fixed probability p greater than one half of correctly determining the state of s, an independence condition which says that each agent's determination is mutually independent of all the other m is one agent's determination.

You have an honesty condition that each agent honestly reports their determinations. Then you want to have a large population of agents and then as m approaches infinity, the probability that the majority of agents is correct quickly approaches one. For example you've got about 10,000 photos, each with 51% probability of being correct, that majority has a probability of about 99% of being correct.

So there are several extensions to jury theorem, for example they've done it so they have allowed variations in agent knowledge, they've allowed or correlated determinations, also with not just one statement, multiple statements. But roughly speaking, I mean you can create these background conditions to get these intelligent explosions in different ways but roughly speaking, if you want to be altruistic, you need to ... cause you need to get directed towards a sinister plot, like what action do I need to take to maximize the probability of killing the most puppies over the next ten years, so you don't want that, you want something that's altruistic.

And then you want your agents to roughly be knowledgeable, or [inaudible 00:47:30], for whatever background conditions you need. You want them to be independent for the most part, you can have some dependency sometimes, you want them roughly to be honest and if you want more agents, usually the better.

Now the extension I do in the paper is just a toy model but basically all it does, it resolves two of the problems. Version ones is that it avoids the intransitive cycles that you have with the majority rule cause majority rule is intransitive. So I avoid that problem. And then also, his model is based on a growing population to increase the probability practice to make the [inaudible 00:48:07] total approach one. I say, okay, let's just have a fixed population, a finite population we're dealing with multiple statements. As we go through more statements, their probabilities of correctness improve approaching to one.

But again, it's just a toy model, a proof of concept so that modelers dealing with other situations can adjust the parameters and variables for the object of applications to figure out how they want to get intelligent explosion for their crowdsourcing applications.

But the thing is, one of the key things about this to make, if you're having multiple statements where you kind of keep using the crowdsourcing application, you want it to keep improving its probability practice and it's going to knowledge sustainable in the long run.

So basically let me go quickly over what Condorcet argued, so one thing, first thing is just basically restatement of his jury theorem, second part there is that basically he argued that the rate of scientific discoveries and the rate of pedagogical improvements are related in an accelerating autocatalytic process which on average improves the rate of discoveries which improves the rate of instruction and so on and so forth.

And then if you're close to where you are in the ideal conditions of the jury theorem, then you can approach one of correctness in your judgements and then you can improve the intellectual capacities and other predictions of a technological singulary hypothesis.

Now in his non-technical philosophical writings such as Sketch, his arguments were dedicated to, these are the conditions we need, what kinds of institutions do we need in real life to get us as close as possible to these conditions. So for example one thing was he had a notion about collective ... a better solution to collective action problem. He had a general assumption, I think it's probably incorrect, but he had a general assumption that people want more not correct knowledge rather than less knowledge. So if you are in a situation, we are in the jury theorem conditions, then you can get more and more knowledge that's approaching one probability practice as opposed to if you're by yourself and you're just trying to figure things out by yourself. So you're gonna wanna work with others but being honest in reporting our new voting, altruistic in helping each other so that you can get this [inaudible 00:50:12] effect. And suppose going by your own and not getting this intelligence explosion.

And I think most formal modeling today would say that's just baloney because you can create different intelligent structures, you can have Gibbard-Satterthwaite theorem things like that nature that can kind of say that that model is wrong.

So in terms of institutions he went through, so he thought that if we improve education and provide universal instruction to all human beings, we could get closest conditions of satisfying for altruism, honesty and knowledge. He argued for institutions like voting by mail instead of in terms of larger fitting in assemblies so then you would maximize the fulfillment of independence conditions when people vote. [inaudible 00:50:49] independent manner, at least closer to that.

He also argued for all humans, regardless of race, sex, class, sexual orientation, to have the right to vote. It basically assists to maximize the population of voters.

And the thing is, Condorcet essentially is the father of social choice theory, he's basically the father of crowdsourcing research and also, the father of intelligence explosion hypothesis. Yet, most of his mathematical and technical writings have not been

collected into his collected works and the vast majority of even his most important works have not been completely translated into English.

Thus, researches have wasted 150 years rediscovering Condorcet's discoveries in social choice before they realized they were duplicating his results. We're today, you know sometimes, duplicating his results on intelligence explosion hypothesis 200 years later.

So, what you can do, we know things like google translate and other efforts to translate natural language text could be used to translate Condorcet's writing as training data. So if you're doing your research on something like that, we could use it as training data and at least get some translations and maybe not rediscover the wheel so often.

All right? Cool, I'm done. Out. Yeah, go ahead.

Speaker 11: I think one of the most contentious pieces of this is something that you kind of [inaudible 00:52:12] most quickly which is the independence hypothesis.

Mahendra: Oh yeah, that's probably not true. Oh yeah, I know.

Speaker 11: But I mean, especially in the contemporary area where we're all obsessed with social networking and these networking connections and if the intelligence explosion of crowdsourcing applications is a dependent on people being able to vote independently, I mean, how do you begin to ... it certainly makes the statement of the problem far more complicated.

Mahendra: Oh, yeah, that it's private.

Speaker 11: How do you deal with that?

Mahendra: So, you know, I don't think it's simple for one thing, but there have been like journalisations of the theorem that use correlated votes instead of independent voters so I mean, some of that goes into technical details and trying to figure out ... I think the idea is like figure out what situation you're in, then figure out maybe an ideal structure of how you're gonna get an intelligence explosion and then be like, all right, so what can I do, what institutions, what things can I do to tweak our situations so we get close to those conditions so we can get that kind of intelligence explosion. That's kind of the framework I'm thinking about.

Cause I mean, I think you're right. Independence assumption by itself is very very strong, it's very difficult to ...

Yeah, go ahead.

Speaker 12: I'll just continue on this. Do you think it's fundamentally not right in this, despite that it does it, because you start off with an independence assumption and we all thought, we get knowledge on some theorem and then you use that knowledge for further steps. So by definition, it's not important anymore.

Mahendra: No, no, no. So what it is-

Speaker 12: So, let me give an example.

Mahendra: Okay, mm-hmm (affirmative)-.

Speaker 12: This has been done in The Netherlands and it was quite contentious. How many muslims do there live in The Netherlands? What comes out, national survey, people say 19 percent, one nine percent. So that is given back a [inaudible 00:54:01], then you can ask more questions, you can make more theorems, actual the real number is five percent. Is a real big difference. So if you go from one to create arguments based on this false knowledge which comes from the crowd because it's only true if it's going to infinity and it's not, so all the rest doesn't really work anymore. So basically this is flawed.

Mahendra: So two things that I would say. One is that the back ... first of all, all of human knowledge is like a wide array of things, right. So a few cases of mistakes, the claims he made are that on average or in expectation. A second thing is that like you say, is it that at least in the version in the paper, independence condition is with respect to the ... and I admit it, it's a toy model, but the independence conditions with respect to terminations on that particular statement. So they don't have to necessarily be independent of other statements that might be brought up.

But like I said before, there are versions of the jury theorem where you'll have a correlated determinations and they exist in literature and you know, you can go in detail later.