ERC Proof of Concept Grant 2017

Annex 1 to the Grant Agreement
(Description of the Action)
Part B

Action Acronym: BOTFIND
Action number: 767454
Action Title: BOTFIND: Finding Bots, Detect Harassing Automation, and Restoring Trust in Social Media Civic Engagement
Principal Investigator: Prof Philip Howard
Host Institution: University of Oxford
Additional Beneficiaries (if applicable):

ERC Proof of Concept Grant 2017
Part B
Section 1: The idea – Excellence in Innovation potential (max. 2 pages)

a. Succinct description of the idea to be taken to proof of concept

Identifying social innovation demand. Increasingly, social media platforms have become tools for manipulating public opinion during elections. Political actors make use of technological proxies in the form of proprietary algorithms and semi-automated social actors—political bots—in subtle attempts to manipulate public opinion. Through the ERC COMPROP Consolidator award, researchers have demonstrated that even simple bots (i) effectively keep negative messages and fake news in circulation longer, (ii) target journalists and civil society groups, and (iii) operate with little oversight from social media firms. Such bots have negative consequences both for public trust in technology innovation and for the quality of public deliberation in Europe's democracies. ERC researchers have been able to identify highly automated, politically-manipulative social media accounts post-hoc, and this Proof of Concept project will allow researchers to take what we have learned and produce an online tool that allows the public to evaluate suspicious social media accounts. Most social media platforms are slow to address troll and bot activity, so this innovative tool will put ERC research into public service in Europe—and around the world.

ERC-funded innovation. Our ERC Consolidator Award (COMPROP, 2015-2020) focuses on computational propaganda and elections. Together with Twitter and Google, we have developed and tested some computational theories about the detection of automated social media activity. During several important political events—referenda and elections—we have tracked automated social media activity and exposed its impact on public life. We have consulted widely with the technologists behind several social media platforms, tested our assumptions on large amounts of Twitter data pulled during multiple elections, and our research has been widely disseminated through media outlets across Europe. Our innovative thinking on bot activity has allowed the team to undertake real-time social and computational science. The next step, to allow our ERC-funded innovation to have a public impact, is to build an accessible online tool that allows the public to evaluate computational propaganda that may be coming over their own social networks.

Innovation potential. The ERC-funded COMPROP project supported the basic social and computational science that allowed researchers to track social media automation through large datasets of content. We have already had a major international social impact by making computational propaganda and social media manipulation into major news stories in several languages across many countries in Europe. There are several important elections coming in 2017 and 2018, so now is the time to translate our theory and research into an innovative public tool.

Program deployment. In the first quarter of this project, the project will build a basic website, mobile application and plugin that allows users to run the diagnostics that we have been using in our large scale post-hoc analysis, and offer an account of the use of automation, fake news, or suspicious news links over Twitter. In the second quarter, we will open the system for limited testing with collaborators and craft the multilingual documentation that teaches users how to interpret findings. The third and fourth quarters will be dedicated to dissemination and fine-tuning the system. Public dissemination—getting the tool used—is key to the success of the instrument and vital to raising citizens' sophistication with social media.

Proof of concept processes. The proofing process involves making and field testing a user-friendly version of the COMPROP research process. The first step is to turn the algorithmic analytical steps we have been doing by hand, with multiple open source applications, on large amounts of data, into a single analytical system that can run automatically for average users. The second step is to build our automated evaluation system into a user-friendly online interface. The third step is to provide open access and have users give feedback. This will allow us to establish tool viability, identify technical issues early, and shape advertising.

Social benefit. This project aims for civic impact and social benefit, more than commercial impact. Public trust in political institutions and social media platforms is declining across Europe. By exploiting the innovation potential of ERC research, we seek to raise the informational sophistication of European voters, providing them with a tool for evaluating public discourse. This online system of evaluating suspicious social media accounts will help journalists, public policy makers, civil society leaders, politicians, and members of the interested public evaluate other users. By helping to identify trolls and bots, this tool may help restore public trust in technology and the process of modern deliberative democracy.

b. Demonstration of Innovation Potential

Program model. During elections and political crises, a growing number of social media users encounter various forms of automated political content, bot-driven activity, or fake news. COMPROP has demonstrated ways of collecting metrics for the kinds of content that tend to degrade civic conversations: social media

content generated by highly automated accounts; negative ideological valences; content that includes links to untrustworthy news sites. The potential innovation here is in turning what we have learned in an academic context into a tool that allows members of the public to run their own diagnostic tests on accounts they find in their own immediate social media networks. In other words, the discoveries we have made with machine learning and large scale analysis of millions of tweets—if accessibly packaged for a wide population of users—could allow average users to evaluate the bias and trustworthiness of news and information coming to them over their own social media networks. Thus there is potential for large-scale expansion at low cost, and maximising the value of excellent ERC-funded research.

Demonstrated potential for effectiveness. In our COMPROP study of social media activity around the Brexit Referendum, we assembled a data set of more than 1.5 million Tweets collected June 5-12, 2016, using a combination of pro-leave, pro-remain and neutral hashtags to collect the data. This sampling strategy yielded 313,832 distinct Twitter user accounts, summarized in Table 1. We found that political bots had a small but strategic role in the referendum conversations: (1) the family of hashtags associated with the argument for leaving the EU dominates, (2) different perspectives on the issue utilize different levels of automation, and (3) less than 1 percent of sampled accounts generate almost a third of all the messages.[1]

Table 1: Computational Propaganda on Twitter During Brexit Referendum, By Account Type

	All Tweets		Heavy Automation		Disclosed Bots	
	N	%	N	%	N	%
Exclusively StrongerIn (number of tweets using one or more of only #strongerin hashtags)	186,279	14.6	28,075	15.1	196	0.1
Exclusively Brexit (number of tweets that used one or more of only #brexit hashtags)	662,745	51.8	97,431	14.7	842	0.1
Exclusively Neutral (number of tweets that used one or more of only Neutral hashtags)	234,170	18.3	13,436	5.7	253	0.1
Mixed, Brexit-Neutral	69,322	5.4	11,667	16.8	72	0.1
Mixed, StrongerIn-Neutral	35,412	2.8	5,099	14.4	44	0.1
Mixed, Brexit-StrongerIn	49,556	3.9	9,735	19.6	89	0.2
Mixed, Brexit-StrongerIn-Neutral	40,926	3.2	13,640	33.3	35	0.1
Total	1,278,410	100.0	179,083	14.0	1,531	0.8

Source: Author's calculations based on sample 06/5-12/2016. "Heavy Automation" = more than 50 tweets/day, "Disclosed Bots" = the account self-identified as a bot or used a known bot launching platform.

Some of the most highly automated accounts then began generating content about the United States. In a larger test of our assumptions about bots we captured some 50m tweets over the course of three Presidential debates and the week of the election. During the election itself, we found that that political bot activity reached an all-time high for the 2016 campaign. (1) Not only did the pace of highly automated pro-Trump activity increase over time, but the gap between highly automated pro-Trump and pro-Clinton activity widened from 4:1 during the first debate to 5:1 by election day. (2) The use of automated accounts was deliberate and strategic throughout the election, most clearly with pro-Trump campaigners and programmers who carefully adjusted the timing of content production during the debates, strategically colonized pro-Clinton hashtags, and then disabled activities after Election Day.[2] Over time we refined our research process: building in redundant servers for ensuring data capture; standardizing the metrics we use for evaluating the impact of bot activity on public conversation; polishing our language for describing our methods. More outcomes will be available as we study the French election in the spring of 2017, but the existing results from ERC COMPROP research demonstrate strong potential for innovation and expansion.

Demonstrated potential for scale. Initial proof-of-concept has been established through repeated iterations of our assumption tests, over multiple countries. Our analysis has been working with data collected from real-world conditions, though it is post-hoc analysis completed after political events. The goal for this Proof of Concept project is to provide the tool for civic use *during* public policy debates, national elections, or political crises. Thus, the potential for scalability has already been established.

[1] Philip N. Howard and Bence Kollanyi, "Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum," *arXiv:1606.06356 [Physics]*, June 20, 2016, http://arxiv.org/abs/1606.06356.
[2] Bence Kollanyi, Philip N. Howard, and Samuel C. Woolley, "Bots and Automation over Twitter during the U.S. Election," Data Memo (Oxford, UK: Project on Computational Propaganda, November 17, 2016), http://www.politicalbots.org.

Section 2 – Expected Impact (max. 2 pages):

a. Economic and/or societal benefits

The societal and economic benefits of raising public trust in social media, transparency in political communication and information quality in civic debate are massive. Computational propaganda degrades the public sphere, causing significant amounts of misinformation and negative messages to stay in circulation at sensitive moments. In particular, highly automated social media accounts are used to present the appearance of consensus or organized opposition where there is none. This has proven to be particularly harmful on issues where the scientific consensus (often arrived at through the support of public science agencies like the ERC) on climate change, immigration, or economic policy is strong. The economic cost of poorly made public policy choices are difficult to estimate, but there is strong evidence that dubious information, spread over social media, misinformed UK voters during the Brexit referendum and US voters during the recent Presidential election. For many citizens, the result is not only diminished trust in our institutions of governance, but in the use value of technologies like innovative social media platforms. By reducing the ability of computational propaganda to misinform citizens, we can increase public capacity to deliberate, express their policy preferences after reviewing high quality information, and elect officials who have evidence-based political platforms. The BOTFIND program has the potential to prevent users from being influenced by highly automated accounts promoting fake news and misinformation. This will be especially true for citizens in countries with major elections or referenda, or citizens in countries involved in a national security crisis or complex humanitarian disaster. A Proof of Concept grant would allow the development of a tool for public use as well as explanatory materials to support journalists, public policy makers, public intellectuals and the interested public learn about political events.

b. Commercialisation process and/or any other exploitation process

For the initial 18-month implementation plan, the Oxford University-based team will not only build and field test the tool, it will also be actively involved in dissemination in multiple European countries having elections in during the year. Active dissemination means meeting with policy makers, civil society groups, journalists, and the interested public to demonstrate the tool. The team will work closely with journalists to raise awareness about the problem of computational propaganda in national newspapers, and provide multilingual explanations that allow European citizens to interpret what they learn from the tool. With this Proof of Concept project, free commercialization means widely advertising the tool as a partial solution to the problems of fake news and social media manipulation. Indeed, the open commercialization process does not require special capital expenditure, other than travel for team members to hold workshops and training meetings in major European capitals, and the maintenance of a website. It is also important to ensure clear, multilingual guidance on interpreting the findings of an individual bot detector. The scientific papers that showcase research using the social and computational assumptions will appear on the website along with information that explains a) what a bot or highly automated social media account is; b) what each of the relevant metrics mean; c) how users should evaluate new sources shared by such accounts; d) what users can do to minimize the influence of such accounts on their social media networks; e) ways of raising trust and improving deliberation over innovative new technology platforms for social networking.

c. Proposed plans for :

***Competitive analysis*.** There are currently only a few other bot detector applications available, and they have mixed record because they generate many "false positive" alerts. The technology firms in the field of social media platforms tend not to be proactive in identifying highly automated accounts that spread false news. In response to some of the negative press coverage they have received in many national news outlets, they have begun programs to crowdsource for news detection and raise user awareness. But platforms like Twitter and Facebook usually only act on highly automated accounts when they have been flagged by users, and most users are unaware of how such accounts have an impact on their social networks. It is precisely this lack of market competition and leadership from technology firms that has led to such international attention to computational propaganda. Scoping has identified two competing programs, the Truthy Bot or Not program from the NSF/Indiana University and the BotAlert program from the University of New Mexico. Both are flawed in different ways: the first regularly produces false positives and negatives, doesn't make use of obvious manual heuristics like account name, only operates on Twitter, was only trained on English commercial spam in 2013, and cannot distinguish between organizational accounts and bots. The second provides only a poorly designed browser interface, offers little explanations to users, and does not incorporate diverse measures of bot detection. In sum, this Proof of Concept project represents a monopoly in terms of scientifically proven research in this domain. This project will use the latest data on automation

over social media platforms, the latest theory on bot detection, and present a user-friendly analytical tool with multilingual documentation.

Testing, technical reports. Testing the assumptions of our bot detection method is a process that has already made significant headway through the basic science work of ERC COMPROP. The field and usability testing of this particular public implementation will begin in the second stage of the project, and will involve having collaborators in government, industry and civil society use and give feedback on the system. This will allow us to make adaptations in the presentation of output before the tool is widely disseminated. Many of the project's methodology articles will serve as technical manuals for the site, but every effort will be made provide user-accessible, multilingual explanations and interpretive guides. The overall effectiveness of the project will be evaluated through a combination of social media usability metrics from people living in Europe, news mentions, and counts of other sites that cross link to our tool. This project will have three ways of measuring success and impact over the course of the project. (1) The first metric of success will come from website usability statistics—large numbers of users from a diverse range of countries will indicate impact. (2) The second metric of success will come from news media mentions: good coverage from national media outlets during elections will evidence a growing ability of journalists to report on technology related issues and rising levels of public exposure to the problem of computational propaganda in deliberation. (3) The third measure of success will be tracking of the number of user accounts that get reported. Indeed, this Proof of Concept project will yield additional data on the behavior of automated social media accounts through a live list of active, suspicious algorithms. The list will be of use to researchers and can be provided to industry to help technology innovators design more trustworthy computing systems.

IPR position and strategy. The application will be offered under creative commons licensing, as will the training materials and the scientific papers provided as background materials to users. More importantly, the code programming API will be shared through Github so that it can be critiqued and improved by a large network of sophisticated users. The other advantage of providing open access rights is that news organizations and civil society partners will be able to implement the bot checking algorithm on their own websites, or additional mobile apps, or through other platforms.

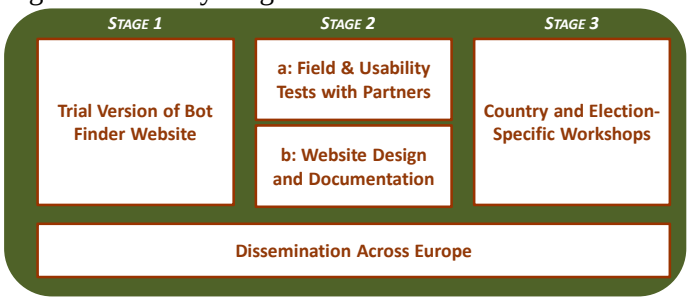
Industry partner(s)/ societal organisation(s) / potential "end users" contacts. Collaboration with NGOs and governments requires sensitive cooperation. It is essential that these partnerships are conducted by project staff with experience in communicating with a range of political and civic actors. A key aspect of establishing innovation potential will be ensuring successful and safe dissemination meetings with relevant country opinion leaders. The management plan identifies some of the civil society groups, government agencies, and news organizations that the principal investigator has ties to and would involve in dissemination. This Proof of Concept project will also develop a wider network of policy and implementation collaborations. Working with the new non-profit NGOs and with partner agencies and governments, we will identify potential new partners in countries that are not having major elections or countries that we cannot visit during the timeframe of this project. We will establish effective communication strategies and systems of providing support to new implementing partners across Europe, such as working with journalists from language specific news agencies and providing virtual training sessions.

The partners for the early stages of this project involving Germany and the upcoming German Election have already been identified for Activity Stages 1 and 2a. Colleagues and collaborators at Der Spiegel are interested in using the tool for their own investigative reporting and in understanding what we find with our research on Germany. Collaborators at Transparency International and the Tactical Technology Collective—both based in Berlin—can help with dissemination by using their extensive civil society networks to drive traffic to our application. The thinktank Stiftung Neue Verantwortung in Berlin can host our workshop for Berlin-based journalists, civil society leaders, policy makers and the interested public. These forms of in-kind support are made possible through existing relationships with the ERC COMPROP team leadership, and do not require subcontracts or other forms of resource transfer.

Section 3: The proof of concept plan (max 2 pages)

a. Plan of the activities. This 18-month project will be conducted by a small, focused Proof of Concept team working alongside the ERC research study team and beginning July 1 2017. The stages involve building the trial version of the website, producing a well-designed, cross platform user interface and field testing it, and then running in-country workshops at select locations around Europe to encourage use. For the duration of the project, the team will be actively engaged with journalists, civil society actors, and public policy makers about the issues of social media algorithms, fake news, and bots in public life (Figure 1).

Figure 1: Activity Stages for BOTFIND



b. Project-management plan. This project will take place alongside the third year of the ERC COMPROP study (alongside follow-up testing, data analysis and publishing). This will allow joint working between researchers and the proof-of-concept team, and capacity-building for public policy makers, civil society leaders, journalists, and the interested public. Thus, Professor Philip N. Howard will lead the project, with a small, specialized Proof of Concept team led by a Postdoctoral / Graduate Innovation Officer, with extensive experience in a) research on social media networks; b) the design and usability testing of online tools; c) engagement with the public on issues of information politics and policy. This post will be supported by a Technical Development Officer, who will support toolkit design, develop IT-based systems and training materials. Rigorous usability testing is essential for the success of the service, so this budget includes both hardware purchases and mobile telephony / data subscription services. The Proof of Concept team will also establish a clear strategy for international scalability. So far, we have responded to requests from governments, journalists and civil society groups on an ad hoc basis, as international crises evolve and time permits. This Proof of Concept team will proactively identify next topics and countries where high levels of bot activity are likely and develop a systematic way of responding to requests for assistance through a purpose built, publically accessible evaluation tool and a suitable program of training materials.

The project plan will be to actively work to improve public understanding of automation, political manipulation, and junk science over social media networks in advance of major democratic exercises. Thus, the dissemination strategy will involve short trips for two personnel to Berlin in October 2017 in advance of the German election, Rome in March 2018 of the Italian election, and visit to Brussels or one other European capital. This third destination will be selected in late 2017 so that we can chose a country election where there will be maximum utility and impact. Looking ahead at the electoral calendar, Sweden, Finland and Hungary are possibilities, but this final destination should be shaped by current events as they are towards the end of the project.

The risks associated with this project are minimal and are greatly outweighed by the potential benefits. While every effort will be made to keep the website secure, it is possible that outsiders will attempt to hack or launch a denial of service attack to change results or make the site inaccessible. It is possible that third party digital infrastructure make track or block use of the site within national borders. The site will not need to collect personally identifiable information from users and will not require user registration, though it will preserve lists of searched accounts for research analysis. Some reasonable security steps, including the provision of https access, which have been budgeted for, will lower our attack surface.

Table 2 outlines the activity stages of the BOTFIND project. Prior to the arrival of resources we will prepare our application for approval to proceed under the university's ethics guidelines and calendar the week-to-week software development stages, milestones, and deadlines. Activity Stage 1involves preparing the first version of the user account query system, sharing code with other collaborators and do a soft launch in time for the German election. Activity Stage 2 has two concurrent components: we must field the instrument with our German contacts and collect some basic usability data from users and make basic improvements in the design of the interface. So while the instrument is being fielded we will compose accessible instructions and interpretive guides in several languages, based on the linguistic skillset in the team—we currently have English, French and German covered. Writing these public-facing documents is the team task while our collaborators take several weeks to play with the tool and send us feedback. This will also be an intensive design stage where we work on the visual presentation of the tool. Stage 3 involves wider dissemination, both through a concerted effort to engage with journalists and through specialized workshops in European capital cities where we can maximize exposure to journalists, policy makers, and the

interested public. The BOTFIND team will forgo scholarly publishing efforts for much of this production schedule, but after the intensive dissemination activities of Stage 3 we will turn to writing for a scholarly audience on topics such as producing public-facing research tools and the process of bot detection. Because the BOTFIND Proof of Concept grant will run during the larger ERC COMPROP grant, the PI will continue dissemination activities well beyond the life of BOTFIND.

Table 2: BOTFIND Project Management Plan	2017				2018			
Project Activity Stages	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall
Prior to Research								
1. Finalize approval from human subjects committee								
2. Plan software development stages, milestones, and deadlines								
Activity Stage 1: Trial Version of Bot Finder Website								
1. Turn COMPROP bot identification practises into user query system								
2. Crowd source code on Github for proofing								
3. Soft release in time for German elections in October 2017								
Activity Stage 2a: Field and Usability Tests with Invited Partners								
1. Work with German-based partners to test the tool								
2. Refine and revise based on feedback								
3. Get to a stable version that can be widely disseminated								
Activity Stage 2b: Website Design and Documentation								
1. Design a multi-platform, multi-device interface for PC browser								
2. Design a multi-platform, multi-device interface for mobile devices								
3. Assemble open technical documents and accessibly written FAQs								
Activity Stage 3: Country and Election-Specific Workshops								
1. Germany October 2017								
2. Italy March 2018								
3. Brussels or a third country chosen for impact, closer to events								
Dissemination Across Europe								
1. Conference presentations, media release in London								
2. Produce online training materials, presentations and videos								
3. Remote training for journalists, policy makers, civic leaders								
Post Research								
1. Follow on analysis in scholarly papers								
2. Project closing documents and filings								
3. Additional dissemination activities								

c. Description of the team. The team has developed considerable experience through the ERC COMPROP project in responding to queries from government agencies, working journalists and civil society leaders about trolls, fake news, and bot activity over information infrastructure. *Principle Investigator*: Professor Philip N. Howard will lead at a minimal cost to the grant. He is a professor at Oxford University and has advised a wide range of organizations on the nature and impact of computational propaganda. *Postdoctoral / Graduate Innovation Officer:* Miss Lisa-Maria Neudert has worked on aspects of the ERC COMPROP project and is an ideal candidate for this role. *Technical Development Officer*: , this position will be recruited in line with the University of Oxford's regular recruitment process. Having a diverse team is key to the successful dissemination of this Proof of Concept project. Together, this team forms an ideal nexus of research and innovation skills to work in conjunction with the existing COMPROP team.

Note that the existing highly trafficked COMPROP website, and the professional contacts with journalists, civil society groups, and policy makers across Europe that have been cultivated during the first two years of the project, are a significant existing resource that will contribute to the success of the project.