Research Statement
Shomir Wilson

Synopsis

The responsibilities that we delegate to computers are ever-increasing in richness and complexity. Necessarily, how we delegate tasks to our computing technology is changing as well. This change is happening with the aid of assistants, agents, and other models that shape sophisticated machine action as directed by human intent. My research in the areas of usable privacy, natural language processing, and artificial intelligence has supported this movement by seeking to answer questions such as the following:

- Users of websites and mobile apps face many challenges to their privacy. The stakes are high for users, who wish to avoid oversharing or misappropriation of their personal data. They are also high for providers of websites and mobile apps, which thrive only with public trust. By helping users to reflect on what they disclose online, is it possible to help them make privacy decisions that align with their professed preferences? Can crowdsourcing, natural language processing, and machine learning bridge the gap between companies' privacy practices and users' understanding of those practices?
- Language technologies rely on the resources we give them, but their efficacy is limited by how well they exploit all the available channels of meaning. One frequently overlooked channel is *metalanguage*, which provides direct, salient information about language in written and spoken contexts. By learning how people refer to document entities, can computers annotate those entities with descriptive tags, context-sensitive summaries, and other valuable information? Can computers acquire information about language by recognizing how people write or speak about language?
- Artificial intelligence has endeavored to solve many domain-specific problems, but *brittleness*—the tendency of an AI system to break when facing the unexpected—remains widespread. Is it possible for AI systems to mimic the introspective resilience that humans exhibit when faced with anomalies or the unanticipated? To what extent can metacognition contribute to the robustness of an AI system?

These threads are unified by a theme that I have named *information introspection*: solving problems with metareasoning (the ability to consider and modify one's reasoning [1]) sourced from computing technologies and from their users. As a career goal, I will add the analysis of embedded meta-level information to the toolbox of techniques that computer science and related fields use to solve problems. By drawing together topics that have had limited contact, I will enrich innovation in each of them. Below I describe my research on three topics, and I include with each of them my plans for future work.

Usable Privacy: Helping People Decide What to Share

Internet users are compelled to share personal information with and communicate over an ever-increasing variety of online services: social networks, retailers, search engines, content providers, and advertisers compete for transactions and, implicitly, trust. The power of sharing and the potential for unintended consequences have made privacy a critical concern. Users' privacy preferences are nuanced and their privacy choices are complex, creating a challenge for all involved parties. This challenge motivates my research in the area of *usable privacy*. I subdivide this work into two parallel threads:

Regret and Privacy in Online Social Networks: It is difficult for users of online social networks (OSNs) to realize the ramifications of sharing decisions, which are a product of both the content they share and the people they share it with. Lapses in judgment can cause a user to deviate from their professed sharing preferences and lead to decisions they regret, with adverse effects on their personal and professional lives [2]. Public figures sometimes provide spectacular examples: they make the news by accidentally oversharing information (including photos) or by posting content that they later regret. Lapses in judgment are familiar most OSN users, through their own errors or from mistakes made by their peers. To study regret in OSNs, my collaborators and I performed a large-scale study [3] of 1.6M deleted tweets collected over a one-week period from a set of 292K Twitter users. Aggregate analysis showed that deleted tweets differ significantly from undeleted tweets on sentiment vocabulary, posting time, and posting location (when available). Many tweets are deleted for reasons unrelated to regret (e.g., typos or

profile management), but the differences confirmed or refuted many of our hypotheses about common scenarios that lead to deletion. Such scenarios may be avoided by effectively informing OSN users about whom they are sharing with and how the content they post relates to their norms. These results motivate my future work to construct effective interventions to help users maintain self-professed posting norms.

Accordingly, efforts to situate OSN users with the reach and self-normativity of their posts are compelling, but it remains challenging to capture users' complex and diverse sharing preferences [4]. One proposed approach for capturing these nuances is the provision of a small number of privacy profiles that express common preferences and provide a basis for configuring individual settings. Within the popular domain of location sharing, I lead a team to perform an in situ user study to determine the effects of privacy profiles on users' sharing choices and their satisfaction with those choices [5]. The results showed that privacy profiles can have a substantial impact on users; for example, users who had access to profiles shared more than users who did not, without an appreciable difference in satisfaction. Thus, designers of privacy settings must carefully consider how simplifying elements are offered to users. The effects of approaches to simplifying privacy choices will be a continuing topic in my research.

The Usable Privacy Policy Project: The status quo approach to notice and choice for online privacy is hostile toward internet users. Websites and apps contain privacy policies that are long and complex, while most users do not read those policies and would struggle to understand them [6]. Poorly informed users have their personal information applied and shared in ways contrary to their wishes, and their lack of understanding stops them from making meaningful privacy choices. Voluntary standards for expressing privacy practices, such as P3P, have produced tepid adoption and mixed results. If internet users are to make meaningful privacy decisions, then privacy policies must become more usable.

My collaborators and I have created The Usable Privacy Policy Project (www.usableprivacy.org) as a new approach to bridging the gap between privacy policies and internet users. This NSF-funded Frontier project is developing a system of crowdsourcing, machine learning, and natural language processing (NLP) to automatically and semi-automatically extract important details from privacy policy text [7]. Early results have shown the feasibility of automatically identifying relevant segments of policy text for crowdworkers to annotate, reducing their workload [8]. One of our goals is to eliminate the need for crowdsourcing in the pipeline, and to do this we are using lightweight semantic parsing to identify roles, actions, and other key entities that characterize a privacy policy. This project is also developing a browser plugin for internet users to understand at a glance the policies of websites they visit. As the NLP and machine learning lead in this project, I expect to lead additional future projects to improve usable privacy for internet users by analyzing documents such as app descriptions and terms of service agreements. Additionally, to increase dialog between the NLP and privacy research communities, I am leading the submission of an "NLP for Privacy" workshop proposal at an upcoming conference.

Natural Language Processing: Discovering How Writers Integrate the Parts of a Text

Research efforts in NLP and language technologies have developed many strategies to make sense of the language that people produce. Widely-discussed approaches to information extraction and semantic parsing can extract information about artifacts of communication, such as terminology, chapters, images, or topics, with nominal precision if given sufficient volumes of data. However, hidden in plain sight is salient information about these artifacts, expressed in phenomena that describe them directly and concisely. The value of such "language about language", termed *metalanguage*, has been noted for NLP, human-computer interaction (HCI), and cognitive science [9]. Research that I led at the University of Maryland and Carnegie Mellon University produced the first viable corpus of *mentioned language*, a specific variety of metalanguage [10], and demonstrated machine learning methods to detect it [11]. The contrast between used language and mentioned language appears in the two sentences below:

(1) The *snow* fell on the wet pavement.

(2) The term *snow* refers to a kind of frozen precipitation.

This research showed that mentioned language frequently occurs in text and is used to accomplish a variety of tasks such as the introduction of new terminology, reporting quotations, emphasis, discussion

of meaning, and clarification of misunderstandings. The next steps in this research are both pure and applied: I am interested in integrating mentioned language into semantic parsing, and I would like to collaborate with other researchers on projects to use mentioned language detection and generation to solve problems in second language learning, dialog systems, and automatic proofreading.

Another common variety of metalanguage consists of references to document entities (*DE references*). Document entities include orthographically structured items (e.g., illustrations, sections, lists) and discourse entities (arguments, suggestions, points). Such references are vital to the interpretation of documents, but they often eschew identifiers such as "Figure 1" for inexplicit references like "in this figure" or "for the exercises below" [12]. My NSF IRFP-funded project at the University of Edinburgh and Carnegie Mellon University has shown the feasibility of an approach to detecting DE references in text by recasting the problem into the classification of word senses, with supportive results from a cross-domain experiment that included documents from Wikipedia, Wikibooks, and website privacy policies [13]. Identifying DE references will enable language technologies to use the information that such references encode, permitting the automatic generation of finely-tuned contextual descriptions of document entities and the creation of tools to aid readers who wish to quickly skim documents for information. My collaborators and I are submitting a grant proposal to the NSF this fall for a project to study how automatically resolving DE references in online course materials can aid student learning.

Artificial Intelligence: Building Resilient Systems Through Metacognition

Artificial intelligence has made great strides toward some of its domain-specific goals, but brittleness is still pervasive: AI systems often fail when faced with the unexpected [14]. My work at the University of Maryland focused on a metacognitive approach to solving AI brittleness. Here, *metacognition* refers to the ability to self-reflect or "think about thinking". The hypothesis of this work was that metacognitive abilities can enable AI systems to cope with unanticipated difficulties that cause their peers to break down. This research produced the metacognitive loop (MCL), a domain-general approach to AI metacognition. At its core, MCL consists of a sequence of interconnected ontologies that contain (respectively) domain-independent indications of anomalies, potential failures, and responses. MCL was tested with a dialog system and a simulation of an autonomous vehicle. It was shown to be capable of "muddling through" obstacles that comparable non-MCL systems could not handle [1].

I intend to bring the results of this research into a future project to develop personalized assistants that help individuals make privacy-related choices when interacting with apps, social networks, and websites. The multifaceted and enduring nature of privacy decisions requires any assistant to be aware of its limitations—unknown factors in decision making and changing preferences, for example—and MCL will serve as a mechanism for the assistant to manage and learn from the unknowns it encounters.

References to My Work (Items below are shortened for brevity; see my CV for more information)

[1] The metacognitive loop and reasoning about anomalies. In *Metareasoning: Thinking About Thinking*, 2010.
[3] Tweets are forever: A large-scale quantitative analysis of deleted tweets. CSCW 2013.
[5] Privacy manipulation and acclimation in a location sharing application. Ubicomp 2013.
[7] Towards usable privacy policies: Semi-automatically extracting data practices from websites'… SOUPS 2014.
[8] Identifying relevant text fragments to help crowdsource privacy policy annotations. AAAI HCOMP 2014.
[10] The creation of a corpus of English metalanguage. ACL 2012.
[11] Toward automatic processing of English metalanguage. IJCNLP 2013.
[12] Determiner-established deixis to communicative artifacts in pedagogical text. ACL 2014.
[13] This table is different: A word-sense based approach to identifying references to document entities. Pending submission for publication; available upon request.
[14] A self-help guide for autonomous systems. *AI Magazine*, Summer 2008.

References to Others' Work

[2] I read my Twitter the next morning and was astonished… Sleeper, M., et al. CHI 2013.
[4] Location privacy in pervasive computing. Beresford, A. and Stajano, F. *Pervasive Comp.* 2(1) 2005.
[6] The cost of reading privacy policies. McDonald, M. and Cranor, L. F. *I/S: J. Law & Policy Info. Soc.* 4(3) 2008.
[9] The use-mention distinction and its importance… Anderson, M., et al. Wkshp. on Sem. & Prag. of Dialog 2002.